МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

М.А. Скулиш,
С.В. Суліма

ОСОБЛИВОСТІ ОБЧИСЛЮВАЛЬНОЇ ІНФРАСТРУКТУРИ ДЛЯ СИСТЕМ КЕРУВАННЯ ТЕЛЕКОМУНІКАЦІЯМИ

Навчальний посібник

Рекомендовано Методичною радою КПІ ім. Ігоря Сікорського
як навчальний посібник для здобувачів ступеня магістра
за освітньою програмою «Інформаційно-комунікаційні технології»
спеціальності 172 Телекомунікації та радіотехніка

Електронне мережне навчальне видання

Київ
КПІ ім. Ігоря Сікорського
2022

Рецензент Кравчук С.О., д.т.н., проф., КПІ імені Ігоря Сікорського
Семенів, А.О., д.т.н., проф.,. Вінницький Національний Технічний Університет

Відповідальний редактор *Скулиш, М.А.,* д.т.н.,с.н.с., КПІ імені Ігоря Сікорського

Гриф надано Гриф надано Методичною радою КПІ ім. Ігоря Сікорського (протокол No 1 від 02.09.2022 р.)

за поданням Вченої ради Навчально-наукового інституту телекомунікаційних систем (протокол No 06 від 12.08.2022 р.)

У навчальному посібнику викладено основні положення методів організації обчислювальної інфраструктури для систем керування телекомунікаціями, що складається з елементів фізичної інформаційно-телекомунікаційної інфраструктури та віртуального обчислювального середовища. У навчальному міститься аналіз систем керування телекомунікація ми, значну увагу приділено задачам організації обчислювальних ресурсів та засобам забезпечення якісного обслуговування кінцевих користувачів мобільних телекомунікаційних мереж. Запропоновані рішення за рахунок систематизації та удосконалення методів керування процесом обслуговування у гетерогенному телекомунікаційному середовищі і адаптивного вибору обчислювальних ресурсів дозволяють забезпечити контрольоване використання ресурсів в умовах різнорідної структури сервісів, підвищення вимог до показників якості обслуговування та неперервного зростанням об'ємів трафіку. Навчальний посібник призначений для здобувачів ступеня магістр за спеціальністю 172 «Телекомунікації та радіотехніка», а також буде корисним для фахівців, які розробляють програмне забезпечення для телекомунікаційних систем.

ЗМІСТ

ВСТУП

Телекомунікаційна мережа оператора мобільного зв'язку сьогодні – це організована система, яка включає в себе спеціальне обладнання, яке обслуговується, спостерігається та керується з операційних дата центрів, де встановлено обчислювальні сервери та відповідне програмне забезпечення, що обслуговує чисельні інформаційні та службові потоки. Сучасні технології SDN, NFV, SDR, CloudRAN та інші стрімко розвиваються, повномасштабне їх впровадження призведе до повної залежності працездатності телекомунікаційної мережі від роботи інформаційно-обчислювального середовища.

Постачальники послуг тепер віртуалізують частини своєї мережі, що суттєво впливає на рішення для аналізу та оцінки параметрів функціонування, які використовуються для забезпечення роботи мережі. Однією з областей розвитку систем забезпечення якості обслуговування абонентів в умовах часткової віртуалізації систем зв'язку є контроль показників якості на всіх етапах надання сервісів абонентам при використанні як телекомунікаційного обладнання так і обчислювальних систем.

Спостерігається інтеграція телекомунікаційних систем (ТКС) та розподіленого обчислювального середовища, в результаті утворюється єдине гетерогенне середовище обслуговування телекомунікаційних сервісів, в якому є можливість контролювати процес обслуговування інформаційних потоків на кожному етапі та забезпечити відповідність високим стандартам якості. У той же час досі немає єдиної концепції, моделей та методів організації контролю та наскрізного керування, які б враховували особливості розгортання мережі дата центрів, обмеження мережевих інтерфейсів, які їх поєднують, особливості логічної структури віртуальної обчислювальної мережі, яка розгортається поверх фізичної інфраструктури, а також систем оркестрування і організації взаємодії всіх елементів та підсистем, які забезпечують процес надання телекомунікаційних послуг. Це призводить до неефективного використання ресурсів, що утворюють гетерогенне середовище обслуговування телекомунікаційних сервісів.

Особливості розвитку інформаційно-телекомунікаційних мереж:

- У зв'язку зі зближенням індустрії інформаційних та комунікаційних технологій, телекомунікаційна інфраструктура все більше використовує хмарні обчислення. Телекомунікаційні оператори надають хмарні сервіси, а також застосовують технології хмарних обчислень для оптимізації своїх телекомунікаційних платформ та систем підтримки (ITU-T M.3371).
- Спостерігається відсутність зворотного зв'язку між якістю послуг, які надаються, та організацією процесів взаємодії у гетерогенному телекомунікаційному середовищі, як наслідок хаотичне завантаження обчислювальних та телекомунікаційних ресурсів, які забезпечують розподілену гетерогенну систему обслуговування.
- Потреба у гнучких моделях та методах керування якістю обслуговування гібридних телекомунікаційних сервісів, які б використовували переваги гетерогенного телекомунікаційного середовища та враховували особливості обчислювальних процесів.

Впровадження технологій програмно керованих мереж потребує впровадження нових моделей теорії масового обслуговування для оцінки параметрів функціонування системи, своєчасного виконання обчислювальних операцій для забезпечення потреб ТКС.

Через відсутність методологічної бази для організації роботи гетерогенного

телекомунікаційного середовища (ГТС), його ресурси використовуються хаотично, задачі оптимізації вирішені частково або локально, що призводить до погіршення контролю та забезпечення показників якості послуг для кінцевих користувачів. Розроблені в роботі моделі та методи є складовими єдиної архітектури контролю та керування ресурсами і потоками на рівні провайдера мобільного зв'язку.

Таким чином, створення і наукове обґрунтування комплексної методології керування процесом обслуговування у гетерогенному телекомунікаційному середовищі з метою підвищення якості процесу обслуговування гібридних телекомунікаційних сервісів є актуальною науково-технічною проблемою.

РОЗДІЛ 1
АНАЛІЗ СУЧАСНОГО СТАНУ ПРОБЛЕМ КЕРУВАННЯ ІНФОРМАЦІЙНО-ТЕЛЕКОМУНІКАЦІЙНИМИ СИСТЕМАМИ

На сьогоднішній день спостерігається стрімкий розвиток інформаційно-телекомунікаційних систем, які використовуються для надання різнорідних сервісів абонентам та пристроям. За останні 10 років суттєво змінилися підходи та концепції організації систем зв'язку. З метою розширення можливостей телекомунікаційних систем, побудованих на основі спеціалізованого апаратного забезпечення, в умовах зміни підходів до організації обчислювальних процесів та інтеграції обчислювальних систем з телекомунікаційними (хмарні обчислення) світова наукова спільнота активно займається розробкою методологічної бази та стандартизацією роботи нових інформаційно-телекомунікаційних систем.

Розвиток систем зв'язку зумовив появу нових складових інформаційно-комунікаційних систем, означення яких наведено у сучасних стандартах та специфікаціях:

NFV – це концепція мережевої архітектури, яка пропонує використовувати технології віртуалізації для представлення цілих класів функцій мережевих вузлів у вигляді складових елементів, які можуть бути об'єднані разом або пов'язаних у ланцюг для створення телекомунікаційних послуг (сервісів).

Гібридний телекомунікаційний сервіс [1] – сервіс, який складається з компонентів телекомунікацій та хмарних сервісів.

Хмарні обчислення (Cloud Computing) [2] – парадигма забезпечення мережевого доступу до масштабованого і гнучкого набору спільно використовуваних фізичних або віртуальних ресурсів з наданням та адмініструванням ресурсів на основі самообслуговування за запитом.

Вузол обслуговування (відповідно до [3] – мережева точка присутності (N-PoP)) – місце, де мережева функція реалізується як Фізична мережева функція (PNF) або Функція віртуальної мережі (VNF).

Мережева функція (NF) [4] – це функціональний блок всередині мережевої інфраструктури, який має чітко визначені зовнішні інтерфейси та чітко визначену функціональну поведінку. На практиці мережева функція є мережевим вузлом або фізичним пристроєм.

Віртуалізовані мережеві функції (VNF) [4]: принцип розділення мережевих функцій на апаратне забезпечення, яке вони запускають, використовуючи абстракцію віртуальних апаратних засобів.

Інфраструктура NFV [4]: сукупність всіх апаратних і програмних компонентів, які створюють середовище, в якому розміщені віртуалізовані мережеві функції (VNF)

Ресурси NFV-Resource (NFV-Res) [4] існують всередині інфраструктури NFV та можуть бути використані для реалізації мережевих функцій або систем, для забезпечення їхньої якості. До них наприклад можна віднести VeCPU – віртуальний процесор – ресурс віртуальної машини.

Компоненти інфраструктури NFV[4]: ресурси обладнання, які не є реплікабельними, але сприймаються як фізичні компоненти виготовлені в умовах виробництва.

Віртуальна машина (VM) [4]: віртуальний обчислювальний простір, який характеризується багатьма фізичними компонентами (процесор, пам'ять/сховище даних, порти/інтерфейси) та генерується за допомогою гіпервізору, який використовує фізичні ресурси та розміщує на них VM. VM можуть виступати хостингом для окремих компонентів віртуальних машин.

Віртуальна мережа (virtual network – VN) [4]: віртуальна мережа містить інформацію про маршрути віртуальних мережевих інтерфейсів VM та фізичних мережевих інтерфейсів, створюючи необхідну зв'язність. Віртуальна мережа обмежена набором доступних мережевих інтерфейсів.

Прикладний програмний компонент [4]: більш загальній термін для частини програмного забезпечення, яке може бути завантажене на віртуальну машину. Віртуальна мережева функція (VNF) є прикладом віртуального прикладного програмного компоненту.

Багатохмарні обчислення – парадигма для взаємодії між двома та більше постачальниками хмарних послуг.

Для уточненого опису середовища, де здійснюється обслуговування гібридних телекомунікаційних сервісів, наведемо визначення гетерогенного телекомунікаційного середовища, яке є організаційно-технічною сукупністю, що складається з каналів зв'язку (фізичних та віртуальних), мережевих вузлів (фізичних та віртуальних) та обчислювального середовища, яке організовано відповідно до парадигми хмарних обчислень, забезпечує функціонування віртуалізованих телекомунікаційних сутностей.

Робота телекомунікаційних систем, які надають послуги зв'язку, сьогодні пов'язана з такими факторами та тенденціями:

1. Зростання об'ємів трафіків в експоненційній прогресії з кожним роком.

2. Мультисервісний потік потребує диференційованого обслуговування, різні вимоги до показників якості обслуговування, різні вимоги до організації процесу передачі, обліку та тарифікації трафіку.

3. Для гарантій заявлених показників якості обслуговування оператору необхідно постійно контролювати показники якості обслуговування абонентського сервісу та вчасно реагувати на зниження показників якості.

Саме тому для підтримки гнучкості телекомунікаційної системи оператору необхідно впроваджувати рішення з використанням програмного забезпечення на різних ділянках системи обслуговування.

Відповідно до рекомендацій ITU-T у складних об'єднаних послугах з багатим мультимедійним наповненням використовують різноманітні види інфраструктури електрозв'язку та інформаційних технологій (IT). Такі послуги складаються з компонентів окремих послуг, придбаних у третіх сторін або наданим третім сторонам.

В рекомендації ITU-T Y.3520 визначено необхідність розробки рішень [1], які б забезпечили можливість однакового наскрізного керування (включаючи облік) послугами, які надаються окремо і спільно доменами і платформами різних постачальників хмарних послуг.

Досягти постійного забезпечення хмарними послугами, які належать до різних доменів – це нелегке завдання. Для його вирішення потрібен підхід, який би створював умови і підтримував постійний керований доступ до хмарних послуг.

Для досягнення зазначених вище цілей постачальникам хмарних послуг потрібні структурні основи, архітектура, шаблони проектування та приклади передового досвіду [5].

Для забезпечення можливості швидкої розробки і розгортання складових багатохмарних послуг в галузі електрозв'язку, ефективного надання сучасних послуг кінцевим користувачам, оперативного керування ними необхідні відповідні удосконалені стандартні моделі та методи керування процесом обслуговування у гетерогенному телекомунікаційному середовищі.

1.1 Мережі NGN як основа для 5G. Особливості системи керування

Згідно з визначенням, наведеним в Рекомендації MCE-T Y.2001, мережа наступного покоління (NGN) – це мережа з пакетною комутацією, здатна забезпечити користувачів різноманітними вузькосмуговими і широкосмуговими послугами, включаючи послуги телефонного зв'язку. Вона заснована на широкосмуговій мережі з пакетною технологією транспортування, що забезпечує необхідну якість послуг QoS, в якій функції, пов'язані з наданням послуг, не залежать від технологій транспортування інформації. Мережа NGN дає користувачам необмежений доступ до різноманітних послуг провайдерів і підтримує узагальнену мобільність, яка дозволяє користувачам отримати доступ до послуг у будь-якому місці і в будь-який час.

У рекомендації MCE-T Y.2012 перераховані основні принципи функціональної архітектури NGN:

1. Підтримка багатьох технологій доступу – функціональна архітектура NGN вимагає гнучкої конфігурації, необхідної для підтримки груп технологій доступу.
2. Розподілене керування – використання принципу розподіленої обробки в пакетних мережах і підтримка прозорості розташування для розподілених обчислень.
3. Відкрите керування – мережеві інтерфейси керування мають бути відкритими для підтримки процесів створення нових, зміни існуючих послуг та підтримки засобів забезпечення логіки послуг сторонніх постачальників.
4. Незалежність надання послуг – процес надання послуг має бути розділений між функціями транспортної мережі, яка працює з використанням механізму розподіленого відкритого керування. Це підтримує конкурентне оточення під час розвитку NGN, сприяє прискоренню процесів впровадження нових послуг.

Для реалізації цих функцій в Рекомендації ITU-T Y.2011 [2] запропоновано базову еталонну модель NGN, яка включає два рівні: рівень послуг NGN (service stratum) і рівень транспорту NGN (transport stratum), кожен з яких містить по три площини: користувача, керування та менеджменту (Рис. 1.1).

Рис. 1.1 Базова еталонна модель NGN (із рекомендації МСЕ-Т Y.2011)

Функціональність рівнів базової еталонної моделі NGN наведено на Рис. 1.2 [2] (рекомендації ITU-T Y.2012). На кожному з рівнів використовують декілька функцій, так для надання послуг (прикладних сервісів кінцевим користувачам) застосовують функції підтримки прикладних сервісів і послуг, відповідні керуючі функції. NGN підтримує точку сполучення з функціональною групою прикладних сервісів, так званий інтерфейс прикладних сервісів мережі (application network interface – ANI), який реалізує канал взаємодії та обміну інформацією між прикладними сервісами і елементами мережі NGN. ANI забезпечує ресурси, необхідні для реалізації прикладних сервісів. Транспортний рівень забезпечує послуги IP-з'єднань для користувачів мережі NGN за допомогою функцій керування транспортом, включаючи функції керування мережевими підключеннями NACFs (Network Attachment Control Functions) і функції керування ресурсами та доступом RACFs.

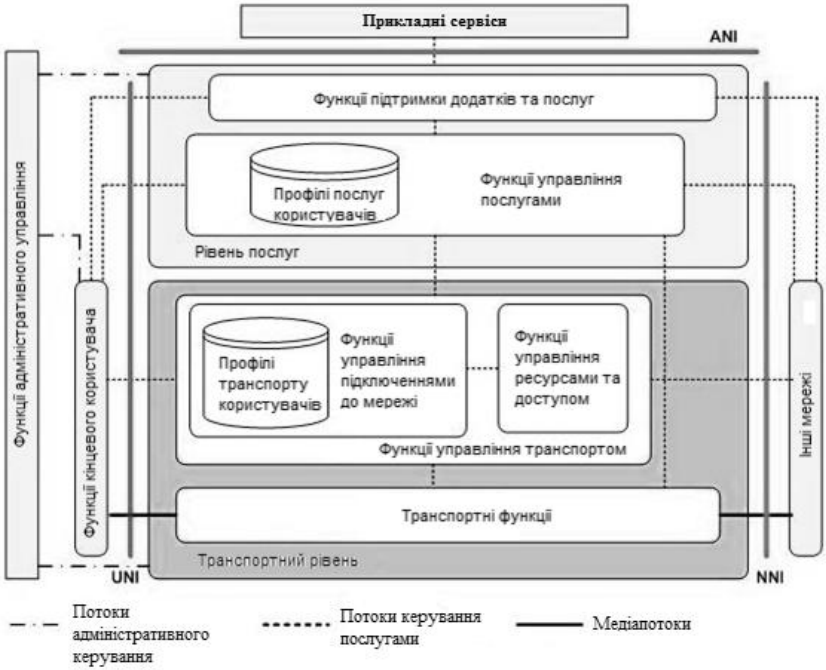


Рис. 1.2 Загальна функціональна архітектура NGN (із рекомендації МСЕ-Т Y.2011)

Відповідно до Рекомендації МСЕ-Т Y.2011 функції транспортного рівня включають безпосередньо транспортні функції і функції керування транспортом.

Транспортні функції (transport functions) забезпечують з'єднання всіх компонентів і фізично розділених функцій всередині NGN. Ці функції підтримують передачу медіаінформації, а також інформації керування (сигналізації) та технічного обслуговування. Транспортні функції включають функції мережі доступу, прикордонні функції, функції транспортного ядра (магістралі) і функції шлюзів.

Функції мережі доступу (access network functions) забезпечують підключення кінцевих користувачів до мережі, а також збір і агрегацію трафіку, що надходить з мережі доступу в транспортну магістраль (ядро). Ці функції також реалізують механізми керування якістю обслуговування QoS, пов'язані безпосередньо з трафіком користувача, включаючи керування буферами, чергами і розкладами, пакетну фільтрацію, класифікацію трафіку, маркування трафіку, визначення політик обслуговування і формування профілю передачі трафіку.

Функції мережі доступу залежать від використовуваної технології доступу, наприклад, вони відрізняються для бездротової технології CDMA та провідної технології доступу xDSL. Залежно від технології, яка використовується для доступу до послуг NGN, мережа доступу включає функції, пов'язані з:

- кабельним доступом;
- доступом за технологіями xDSL;
- бездротовим доступом (наприклад, технології IEEE 802.11 (WiFi), 802.16 (WiMAX), доступ 3G RAN);
- оптичним доступом.

Прикордонні функції (edge functions) використовують для обробки трафіку, який отримують шляхом агрегування трафіку, що надходить з різних мереж доступу і передається в магістральну транспортну мережу. Вони включають функції, пов'язані з підтримкою якості обслуговування QoS і керування трафіком. Прикордонні функції використовують також між магістральними транспортними мережами.

Магістральні транспортні функції (core transport functions) відповідають за гарантовану передачу інформації через транспортну мережу з різним рівнем якості. Вони забезпечують механізми реалізації заданого рівня якості передачі QoS для користувача трафіку, включаючи керування буферами, чергами і розкладом, фільтрацію пакетів, класифікацію, маркування і формування трафіку, контроль дотримання правил обслуговування, керування шлюзами і функції міжмережевих екранів.

Функції шлюзів (gateway functions) забезпечують можливості взаємодії з функціями кінцевих користувачів і/або іншими мережами, включаючи інші типи мереж NGN та ряд існуючих мереж, таких як ТфЗК/ISDN, публічний Інтернет та інші. Функції шлюзів можуть керуватися або безпосередньо функціями рівня керування або через функції керування транспортною мережею.

Функції обробки медіаінформації (media handling functions) забезпечують обробку медіаінформації при наданні послуг, таких як генерація тональних сигналів і перекодування. Ці функції реалізують спеціальними ресурсами обробки медіаінформації на транспортному рівні.

Функції керування транспортною мережею (transport control functions) включають функції керування ресурсами та доступом, функції керування приєднанням до мережі.

Функції керування ресурсами та доступом RACFs діють як арбітр між функціями керування послугами і транспортними функціями для підтримки QoS і пов'язані з керуванням транспортними ресурсами в мережі доступу і в магістральній транспортній мережі. Рішення з керування ґрунтується на інформації про необхідий транспорт, угодах про заданий рівень обслуговування SLA, правилах мережевої політики, пріоритетах послуг та інформації про стан і використання транспортних ресурсів. Функції RACF забезпечують абстрактний підхід до інфраструктури транспортної мережі для функцій керування послугами SCFs і надають сервіс-провайдерам можливість не залежати від мережевої топології, зв'язності, завантаження ресурсів, механізмів/технологій QoS та ін. Функції RACF взаємодіють з функціями SCF і транспортними функціями для різних прикладних програмних компонентів (наприклад, SIP-виклики, потокове відео й ін.), що вимагає керування транспортними ресурсами NGN, включаючи керування QoS, керування NAPT/Firewall і проходження трансляції мережевих адрес на рівні портів NAPT.

Функції керування підключенням до мережі NACFs забезпечують реєстрацію на рівні доступу та ініціалізацію функцій кінцевого користувача для послуг доступу NGN. Ці функції забезпечують транспортний рівень ідентифікацією/авторизацією, керуючи простором IP-адрес в мережі доступу і аутентифікації сесій доступу. Вони також

повідомляють кінцевим користувачам про контактні точки до функцій NGN на рівні послуг. Функції NACF включають транспортний профіль користувача, який зберігатися у вигляді функціональної бази даних, що включає інформацію користувача, а також інші дані керування.

Рівень послуг (service stratum) включає:

- функції керування послугами, включаючи функції профілів послуг користувачів;
- функції підтримки прикладних програмних компонентів і функції підтримки послуг;
- функції кінцевих користувачів;
- функції адміністративного керування.

Функції керування послугами (service control functions) включають керування ресурсами, функції реєстрації, аутентифікації та авторизації для різних сервісів на рівні послуг. Вони також можуть включати функції керування медіаресурсами, такими як спеціалізовані пристрої та шлюзи на сигнальному рівні. Функції керування послугами підтримують профілі послуг користувачів, які є комбінацію інформації користувача та інших даних керування, індивідуальний профіль кожного користувача, такі дані зберігають у функціональних базах даних.

Функції підтримки прикладних програмних компонентів і функції підтримки послуг (application support functions and service support functions) включають функції шлюзів, реєстрації, аутентифікації та авторизації на рівні прикладних програмних компонентів. Ці функції доступні в функціональних групах «прикладні сервіси» і «кінцеві користувачі». Вони працюють спільно з функціями керування послугами для забезпечення кінцевих користувачів і прикладних сервісів необхідними послугами NGN. Через інтерфейс «користувач-мережа» UNI функції підтримки прикладних програмних компонентів і функції підтримки послуг забезпечують точку доступу до функцій кінцевих користувачів. Взаємодія прикладних програмних компонентів з даними функціями здійснюється через точку доступу, реалізовану інтерфейсом «прикладна програма-мережа» ANI.

Функції кінцевих користувачів (end-user functions) не визначають ніяких обмежень на інтерфейси користувача і мережі доступу кінцевих користувачів, які можуть бути з'єднані з мережею доступу NGN. Термінальні пристрої користувачів послуг NGN є будь-якими мобільними або стаціонарними пристроями.

Функції адміністративного керування (management functions) забезпечують можливість керувати мережею NGN для надання послуг із заданим рівнем якості, безпеки та надійності. Ці функції розподіляються децентралізовано по всім функціональним блокам (FE) і вони взаємодіють з функціональними блоками керування мережевими елементами, керування мережею і керування послугами. Функції адміністративного керування використовують на транспортному рівні і рівні послуг і для кожного з рівнів вони реалізують такі завдання:

- керування процесом усунення відмов (fault management);
- керування конфігурацією мережі (configuration management);
- керування розрахунками з користувачами і постачальниками послуг (accounting management);
- контроль продуктивності мережі (performance management);
- забезпечення безпеки роботи мережі (security management).

З метою більш простого розуміння принципів побудови мереж наступного покоління в більшості публікацій з NGN наводиться узагальнена 4-х рівнева архітектура NGN, в якій виділяються такі рівні (Рис. 1.3):

Рис. 1.3 Чотирьохшарова модель NGN

- рівень доступу, який містить мережу абонентського доступу до транспортної пакетної мережі;
- транспортний рівень, який включає магістральну пакетну мережу (мережу, побудовану на базі протоколів пакетної комутації IP або ATM, на сьогоднішній день найчастіше на базі технології MPLS та протоколу IP);
- рівень керування комутацією, включає сукупність функцій з керування усіма процесами обслуговування викликів в телекомунікаційній мережі;
- рівень послуг та експлуатаційного керування, який містить логіку виконання послуг та/або прикладних програмних компонентів, керує цими послугами, має відкриті інтерфейси для використання сторонніми організаціями (для розробки нових сервісів).

Термінальне обладнання рівня доступу не входить до складу мережі NGN. Безпосереднє підключення до мережі можливо тільки для пакетних абонентських терміналів, які працюють з використанням протоколів SIP та H.323.

Концепція NGN спричинила еволюцію систем керування в телекомунікаціях від спеціальних апаратних рішень до програмно-керованих мереж (SDN), які здійснюють контроль та керування відповідним обладнанням гетерогенних інформаційно-телекомунікаційних систем.

1.2 Система організації хмарних обчислень. Аналіз стандартів для забезпечення керування телекомунікаційними системами

Хмарні обчислення (Cloud Computing) [ITU-T Y.3500]: Парадигма забезпечення мережевого доступу до масштабованого і гнучкого набору спільно використовуваних фізичних або віртуальних ресурсів з наданням та адмініструванням ресурсів на основі самообслуговування за запитом.

Прикладами таких ресурсів є сервери, операційні системи, мережі, програмне забезпечення, програми та обладнання для зберігання даних.

Договір про рівень обслуговування (Service Level Agreement) [ITU-T Y.3500]: Документально оформлена угода між постачальником послуги і споживачем, в якій визначаються послуги та цільові показники обслуговування.

Угода про рівень обслуговування може укладатися також між провайдером послуги і постачальником, внутрішньої групою або споживачем, який виступає в ролі постачальника. Договір про рівень обслуговування може включатись в договір або в документально оформлену угоду іншого типу.

Керування ресурсами (Resource management) [5]: найбільш економічний і ефективний спосіб доступу, контролю, керування, розгортання, планування і зв'язування ресурсів, які надаються постачальниками послуг.

Хмарні прикладні програмні компоненти, які також називаються хмарними робочими навантаженнями, – це прикладні програмні компоненти (тобто програми конкретного цільового призначення), які повинні виконуватися в центрах обробки даних постачальника хмарних послуг для створення екземплярів хмарних послуг і забезпечення їх доступності користувачам. Іншими словами, хмарний прикладний програмний компонент повинен виконуватися для того, щоб надати користувачам одну або кілька хмарних послуг.

Постачальникам хмарних послуг необхідно виробити глибоке розуміння аспектів надання послуг, які відносяться до часу виконання, а також методів керування цими послугами і ресурсами, необхідними для їх надання [5].

У складних об’єднаних послугах з багатим мультимедійним наповненням використовують різноманітні види інфраструктури електрозв'язку та інформаційних технологій (ІТ). Такі послуги складаються з компонентів окремих послуг, придбаних у третіх сторін або наданим третім сторонам

Одна з цілей рекомендації ITU-T Y.3520[5] полягає в тому, щоб забезпечити можливість однакового наскрізного керування (включаючи облік) послугами, які надаються окремо і спільно доменами і платформами різних постачальників хмарних послуг.

Досягти постійного забезпечення хмарними послугами, які належать до різних доменів – це нелегке завдання. Для його вирішення бажано виробити підхід, який би створював умови і підтримував постійний доступ керування до хмарних послуг.

Для досягнення зазначених вище цілей постачальникам хмарних послуг потрібні структурні основи, архітектура, шаблони проектування та приклади передового досвіду [5]. Інтерфейси компонентів окремих послуг не є основним предметом розгляду даної роботи, оскільки фізичні інтерфейси можуть відрізнятися в залежності від реалізації, використовуваних сторонніх технологій і вимог операторів. Для забезпечення можливості швидкої розробки і розгортання складових багатохмарних послуг в галузі

електрозв'язку, а також оперативного керування ними, необхідні відповідні стандартні принципи проектування та базові структури.

1.2.1 Особливості керування сервісами із застосуванням хмарних обчислень

Підвищена складність керування ресурсами, пов'язаними з наданням хмарних послуг, обумовлена двома основними особливостями керування сервісами із застосуванням хмарних обчислень. Перша особливість – це віртуалізація обчислювальних і мережних ресурсів в еталонній архітектурі хмарних обчислень [2]. Друга особливість полягає в тому, що в наданні хмарних послуг все частіше беруть участь домени великої кількості постачальників хмарних послуг, і таке гетерогенне середовище надзвичайно ускладнює наскрізне керування ресурсами.

В цілому керування ресурсами слід розглядати з точки зору керування життєвим циклом хмарного прикладного програмного сервісу. На різних етапах свого життєвого циклу він має піддаватися впливам з боку традиційних бізнес-процесів, пов'язаних з виконанням функцій системи керування, таких як адміністрування, підготовка, конфігурація, забезпечення якості послуг і нарахування плати.

Як показано на Рис. 1.4, в простому випадку, коли прикладний програмний сервіс розміщений в однохмарній обчислювальної системі, вона стає залежною від двох окремих категорій віртуалізованих ресурсів. Пунктирними стрілками позначений активний координований зв'язок, який необхідно підтримувати між ресурсами на кожному рівні.

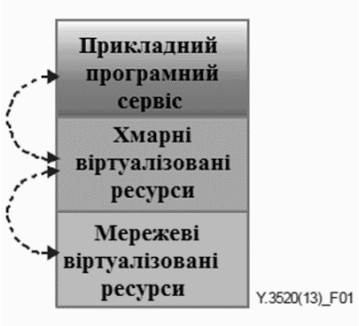


Рис. 1.4 Прикладні програмні сервіси, розміщені в однохмарній обчислювальній системі

Хоча на рис. 1.4 віртуалізовані ресурси діляться на хмарні і мережеві, в хмарних обчисленнях вважається, що всі ресурси розташовуються на одному рівні [4].

Необхідно вирішити питання про те, як засобами існуючих систем керування хмарою забезпечити відстеження логічних і фізичних ресурсів, що реально підходять для певного екземпляру конкретного прикладного програмного сервісу в будь-який заданий момент часу.

Передові практичні методи мають забезпечувати гнучкість хмарного прикладного програмного сервісу з точки зору надання доступу до своїх інтерфейсів керування послугами або до ресурсів. Крім того, вони мають передбачати можливість відкриття доступу до одного або декількох інтерфейсів системи керування, щоб та відстежувала динамічні зміни основних ресурсів, виділених для підтримки роботи керованого хмарного прикладного програмного компонента.

На **Ошибка! Источник ссылки не найден.** показана високорівнева архітектура наскрізного керування багатохмарними ресурсами. Тут зображені віртуальні машини з програмним стеком, який складається з міжплатформених рівнів з серверами прикладних програмних компонентів, що працюють в обраному середовищі виконання, поверх яких виконуються хмарні прикладні програмні компоненти.

Крім того, на **Ошибка! Источник ссылки не найден.** представлені функціональні інтерфейси (FI) і інтерфейси керування послугами (SMI), до яких надають доступ різні хмарні прикладні програмні компоненти, що працюють в центрах обробки даних від безлічі хмар. Необхідну інформацію можна збирати через будь-які інтерфейси SMI, пов'язані з великою кількістю прикладних програмних компонентів, що виконуються в центрах обробки даних від безлічі хмарних послуг. Це дозволяє реалізувати всеосяжне наскрізне керування багатохмарними ресурсами і систему моніторингу.

На **Ошибка! Источник ссылки не найден.** прикладні програмні компоненти, які працюють у віртуальних машинах, можуть бути об’єднаними, розподіленими, побудованими з різних програмних компонентів. Окремий екземпляр віртуальної машини може містити всі програмні компоненти, що відносяться до такого прикладного програмного компонента, або ж тільки деякі з них у разі розподіленого прикладного програмного компонента, що виконується в декількох віртуальних машинах (звідси посилання на прикладні програми або компоненти на **Ошибка! Источник ссылки не найден.**).

Концептуальна архітектура, наведена на **Ошибка! Источник ссылки не найден.**, дозволяє створювати функціонально сумісні програми з підтримкою сценаріїв виносу в хмару або гібридних хмарних обчислень.

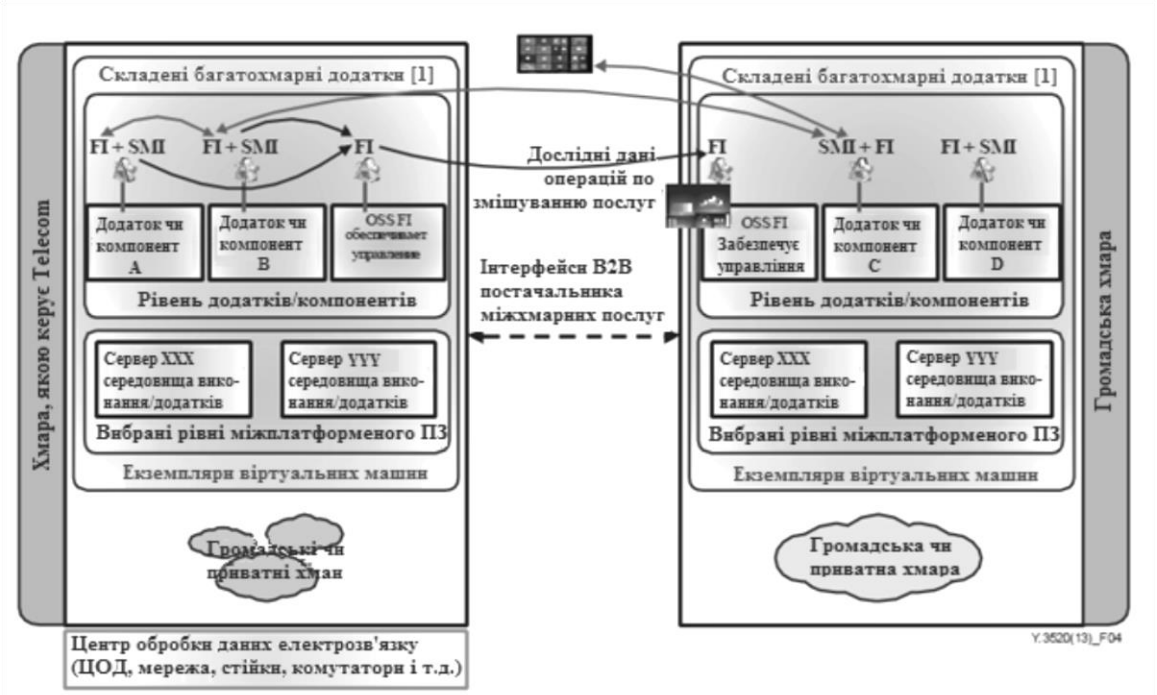


Рис.1.5 Концептуальна архітектура багатохмарного і багатоплатформеного керування хмарними послугами

Між-хмарне обчислення (inter-cloud computing) [2]: парадигма для взаємодії між двома або більше постачальниками хмарних послуг.

Телекомунікаційний сервіс (telecommunication service): те, що адміністрація пропонує своїм клієнтам для задоволення конкретних вимог до електрозв'язку.

Сервіс передачі (bearer service) та телесервіс (teleservice) є різновидами служби телекомунікацій. В майбутньому можуть бути визначені інші види телекомунікаційних послуг.

Інтерфейс керування сервісом (service management interface) [MCE-T M.3070]: інтерфейс, який надає набір можливостей керування, що виникають в рамках хмарного сервісу. З його допомогою здійснюється керування хмарним сервісом.

Постачальник послуг (service provider) [b-ITU-T M.3320]: загальне визначення суб'єкта, який надає телекомунікаційні послуги клієнтам та іншим користувачам на тарифній або контрактній основі. Постачальник послуг може здійснювати керування мережею. Постачальник послуг може бути замовником іншого постачальника послуг.

ТС-гібридний сервіс (ITU-T M.3371): сервіс, який складається як з компонентів телекомунікацій, так і хмарних сервісів.

Як важлива частина керування телекомунікаціями, керування сервісом реалізує всі функції, необхідні для роботи комунікаційних та інформаційних послуг, що надаються клієнтам. Сюди входить виконання послуг, надання послуг та їх оплата протягом всього життєвого циклу послуги:

– *виконання послуг*: виконання ресурсних можливостей та вимог до якості обслуговування клієнтів, підтримка готовності ресурсного та сервісного потенціалу, забезпечення функціональності та інтерфейсу, необхідного клієнтам для доступу та споживання послуг;

– *забезпечення послуг*: підтримка функціональності та процесу надання послуг включає в себе керування рівнем обслуговування, керування інцидентами та проблемами, сервісний моніторинг та керування звітністю;

– *білінг послуг*: підтримка функції та процесу оплати послуг.

У зв'язку з наближенням індустрії інформаційних та комунікаційних технологій, у сфері телекомунікаційної інфраструктури все більше використовують хмарні обчислення. Телекомунікаційні оператори надають хмарні сервіси, а також застосовують технології хмарних обчислень для оптимізації своїх телекомунікаційних платформ та систем підтримки (ITU-T M.3371).

[ITU-T M.3070] визначає загальну модель, засновану на інтерфейсі керування послугами (SMI), для керування хмарними обчисленнями з кінця в кінець (end-to-end E2E), яка описана на Рис. 1.6. У цій моделі система керування телекомунікаціями, яка працює у хмарі, може керувати хмарними об'єктами через інтерфейс керування «I1», який може відповідати набору SMI.

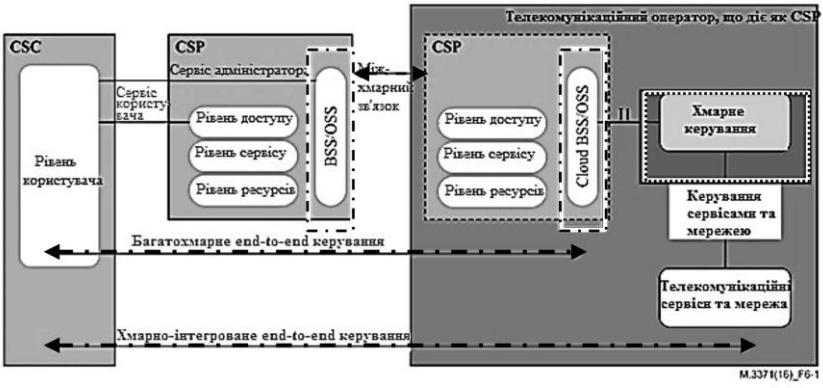


Рис. 1.6 Спільна модель керування хмарними обчисленнями E2E, з оператором зв'язку, що виступає в ролі CSP (після [МСЕ-Т M.3070])

Система керування телекомунікаціями, що працює у хмарі, може послідовно керувати телекомунікаційними та хмарними сервісами. Незважаючи на те, що поділ на рівні керування може бути таким самим, як в класичному телекомунікаційному менеджменті, з впровадженням хмарних обчислень існують ще нові вимоги до керування хмарними телекомунікаціями. Ця рекомендація стосується вимог комплексного керування послугами в хмарній системі керування телекомунікаціями.

Співвідношення між керуванням хмарними сервісами та керуванням сервісами в хмарних телекомунікаціях також показано на Рис. 1.6. Керування хмарними службами знаходиться в системі підтримки бізнесу (BSS)/системі підтримки операцій (OSS) (позначено штрих-пунктирним прямокутником на Рис. 1.6), що належить постачальнику хмарних сервісів (CSP) а керування службами у хмарних телекомунікаціях виконує інформаційна система керування телекомунікаціями (позначена пунктирним прямокутником на Рис. 1.6), що належить оператору телекомунікацій. Ще одним важливим аспектом керування хмарними сервісами є керування життєвим циклом хмарних сервісів (див. [ITU-T Y.3522]), що також знаходиться в BSS/OSS, і належить CSP.

1.2.2 Загальні вимоги до керування сервісами у гетерогенному інформаційно-телекомунікаційному середовищі

Відмінність хмарної системи керування телекомунікаціями від традиційної полягає у підтримці комплексного керування телекомунікаційними сервісами, хмарними сервісами та ТС-гібридними сервісами, які одночасно складаються з телекомунікаційних і хмарних сервісів.

До хмарної системи керування телекомунікаціями висуваються такі вимоги:

– керування послугами повинно забезпечувати керування функціональністю каталогів сервісів та інвентаризації сервісів для забезпечення телекомунікаційних послуг, хмарного сервісу та ТС-гібридної служби;

– керування послугами повинно підтримувати модифікацію та припинення (на замовлення та автоматично) надання телекомунікаційних послуг, хмарного сервісу та ТС-гібридної послуги;

– керування сервісами повинно підтримувати забезпечення якості E2E у сфері телекомунікаційних послуг, хмарного сервісу та ТС-гібридної послуги, а також забезпечувати високий рівень надійності та доступності, відповідно до угоди про рівень обслуговування (SLA);

– керування сервісами повинно підтримувати тарифікацію телекомунікаційних послуг, хмарного сервісу та ТС-гібридних послуг, відповідно до часу використання, пропускної здатності, ресурсів та будь-якої їх комбінації;

– керування сервісами повинно забезпечувати моніторинг, аудит та надання звітності для телекомунікаційних послуг, хмарного сервісу та ТС-гібридних сервісів з метою оцінки та забезпечення якості надання послуг.

Всі перелічені вище вимоги відносяться як до послуг, що надаються єдиним CSP, так і до міжхмарного провайдера.

Функціональна програмна платформа для керування послугами в міжхмарній системі керування

Високорівнева організація функцій керування сервісами в хмарній системі керування телекомунікаціями складається з керування сервісним каталогом, інвентаризацією послуг (service inventory), тестуванням, замовленням, якістю надання послуг, ефективністю обслуговування та оцінки послуг/дисконтуванням.

Рис. 1.7 Функціональна програмна платформа для керування послугами в міжхмарній системі керування

На Рис. 1.7 зображено функціональну основу для керування сервісами в хмарній системі керування телекомунікаціями:

Функції наступні:

– керування каталогом сервісів: надає можливості для створення та розробки нових послуг, визначення їх характеристик, керування комплексними правилами, забезпечує підтримку компонування послуг та керування їх зв'язками і залежностями;

– керування інвентаризацією послуг: надає можливості для зберігання та керування елементами послуг та їх атрибутами. Функціональна платформа також зберігає й керує взаємодією сервісів, тобто відображенням послуг одних послуг для інших та/або їх компонентів;

– керування тестуванням послуг: надає можливості для забезпечення роботи служб належним чином. У процесі виконання, тестування служби відповідає за те, щоб зазначена служба працювала так, як повинна за замовчуванням. У процесі надання послуг, тестування відповідає за розпізнавання проблем та недоліків;

– керування активністю послуг: надає можливості для керування життєвим циклом E2E запиту на обслуговування. Вона охоплює перевірку наявності послуг, а також порядок надходження запитів на обслуговування. Сюди також входить порядок розбиття послуг та/або продуктів на частини, а також відстеження їх обслуговування, процесів активації та тестування;

– керування сервісними проблемами: надає можливості для керування проблемами, що впливають на роботу з клієнтом, а також мережевими проблемами/недоліками та їх ефективне вирішення;

– керування якістю послуг: надає можливості для моніторингу та керування рівнями обслуговування. Показники якості надання послуг збираються та порівнюються з еталонними, а висновки надаються зацікавленим сторонам;

– керування ефективністю обслуговування: надає можливості для моніторингу, аналізу та звітування про ефективність надання послуг E2E. Дана функція повинна давати можливості для аналізу E2E, щоб забезпечити правильне функціонування кожної служби, а також історичний огляд;

– керування оцінкою послуг/дисконтуванням: надає можливості для забезпечення отримання клієнтом рахунку, з відображенням всіх оплачуваних подій, що надаються постачальником послуг, відповідно до ділових взаємовідносин між ними.

1.2.3 Функціональні вимоги до керування сервісами в хмарній системі керування телекомунікаціями

Далі наведені функціональні вимоги до керування сервісами в хмарній системі керування телекомунікаціями.

Керування каталогом послуг

У хмарній системі керування телекомунікаціями вимоги керування каталогом послуг наступні:

- керування каталогом послуг повинно надавати функцій для керування інформацією про хмарні послуги. Потрібно, щоб керування каталогом послуг забезпечувало функції для керування інформацією ТС-гібридних послуг.

Керування інвентаризацією послуг

У хмарній системі керування телекомунікаціями вимоги до інвентаризації послуг наступні:

- керування інвентаризацією сервісів повинно забезпечувати функції для керування атрибутами хмарного сервісу;
- керування інвентаризацією сервісів повинно забезпечувати функції для відстеження взаємозв'язку між хмарним сервісом або гібридною службою ТС і ресурсами.

Керування активацією послуг

У хмарній системі керування телекомунікаціями вимоги до активації послуг наступні:

- надавати функції для перевірки наявності хмарних сервісів, а також запитів на замовлення послуг;
- надавати функції для керування розкладом замовлення послуг та продуктів;
- надавати функції для відстеження процесу підготовки активації послуги.

Керування сервісними проблемами

У хмарній системі керування телекомунікаціями вимоги щодо керування сервісним проблемами наступні:

- надавати функції для керування запитами про скарги з боку клієнта, а також повідомленнями від системи моніторингу про несправність мережі;
- надавати функції для розподілення помилок між хмарною та телекомунікаційною мережею;
- мати цілісне уявлення про конфігурацію та взаємозв'язок між різними мережевих рівнями для забезпечення аналізу первинної причини виникнення проблеми.

Керування ефективністю надання послуг

У хмарній системі керування телекомунікаціями вимоги щодо керування ефективністю надання послуг наступні:

- забезпечувати моніторинг, аналіз та надання звітності про ключовий показник ефективності (KPI) хмарного сервісу.

Керування тестуванням послуг

У хмарній системі керування телекомунікаціями вимоги щодо керування тестуванням послуг наступні:

- надавати функції для тестування хмарного сервісу як у процесі виконання, так і в процесі підтвердження надання послуги.

Керування якістю послуг

У хмарній системі керування телекомунікаціями вимоги щодо керування якістю надання послуг наступні:

- надавати функції моніторингу, аналізу та звітування про ключовий показник якості (KQI) хмарного сервісу.

Керування оцінкою послуг/дисконтуванням

У хмарній системі керування телекомунікаціями вимоги щодо керування оцінкою послуг/дисконтуванням наступні:

- надавати функції для підтримки оцінки та дисконтування послуг в контексті комплектування та складання сервісів із хмарним обслуговуванням.

Наприклад, за рахунок застосування в керуванні концепції загальних моделей мережі, стає можливим загальне керування різнотипним обладнанням, мережами і

послугами, що використовують загальні інформаційні моделі і стандартні інтерфейси (ITU-T M.3060 / Y2401 (03/2006)).

Керування мережами електрозв'язку направлене на підтримку широкої різноманітності областей керування, що охоплює планування встановлення, експлуатацію, адміністрування, технічне обслуговування та надання мереж і послуг електрозв'язку.

В МСЕ-Т керування ділиться на п'ять широких функціональних областей керування (Рекомендація МСЕ-Т М.3400) FCAPS:

- керування обробкою відмов;
- керування конфігурацією;
- керування обліком;
- керування якістю роботи;
- керування безпекою.

Справжня класифікація інформаційного обміну в структурі керування не залежить від того, яким чином буде використовуватися інформація. Для керування мережами електрозв'язку мережі й послуги повинні розглядатися як сукупності взаємодіючих систем. Бізнес-процеси, описані в Рекомендаціях МСЕ-Т серії М.3050.х, і функціональні області керування FCAPS (Обробка відмов, конфігурація, облік, якість роботи, безпека), описані в Рекомендації МСЕ-Т М.3400, повинні розглядатися як теоретичні побудови, необхідні для мереж і послуг наступних поколінь. Архітектура передбачає таку організацію керування окремими системами, при якій на мережі досягається узгоджений результат.

Завдання в галузі керування мережами наступних поколінь включають в себе:

- мінімізацію посередницької роботи між різними технологіями мереж шляхом зближення підходів до керування і використання інтелектуальної звітності;
- мінімізацію часу реагування системи керування на події в мережі;
- мінімізацію навантаження, створюваного трафіком керування;
- географічно розосереджений контроль над аспектами експлуатації мережі;
- надання механізмів розмежування для мінімізації ризиків у сфері безпеки;
- надання механізмів розмежування для виявлення і запобігання збоїв мережі;
- поліпшення допомоги в обслуговуванні і взаємодії з клієнтами;
- багаторівневе представлення послуг, що дозволяє постачальнику надавати структурні блоки для послуг, а також групувати послуги і їх вплив на архітектуру керування;
- бізнес-процеси, визначені в Рекомендаціях серії М.3050.х, а також їх майбутнє використання в СПП;
- підтримку прикладних програмних компонентів як на єдиній розподіленій обчислювальній платформі, так і розподілених по мережі.

Для подальших досліджень визначені наступні області:

- значення потреби в керуванні наскрізними послугами;
- значення домашніх мереж і обладнання в приміщенні користувача.

У той час як рівень керування обслуговуванням (РУО) відповідає за керування життєвим циклом послуг, а також за доставку і забезпечення екземплярів послуг, рівень керування ресурсами (РУР) відповідає за керування логічного інфраструктурою обслуговування та логічною транспортною інфраструктурою.

Функції, які є частиною функцій керування ресурсами забезпечують відображення інформації, орієнтованої на послуги, яка використовується функціями керування обслуговуванням. Дана інформація буде залежати від ресурсів/технологій, яка використовується ресурсами СПП.

У термінах порівняння зі структурою eTOM Рекомендацій МСЕ-Т серії М.3050.x, можна відобразити схожість, як показано на Рис. 1.8:

Рис. 1.8 13/M.3060/Y.2401 – керування ресурсами

Функція керування ресурсами складається з двох головних підфункцій, з'єднаних в місці поділу архітектури СПП на рівень послуг СПП і транспортний рівень СПП:

- функції керування ресурсом обслуговування;
- функції керування транспортним ресурсом.

Функція керування ресурсом обслуговування надає функціональну можливість керування для нового набору характеристик керування ресурсами, що належать до підтримки шару послуг СПП, таких як керування прикладним програмним компонентом, дані програми, користувачі, дані користувачів, кінцеве обладнання і т. д.

Функція керування транспортним ресурсом надає функціональну можливість для традиційних функцій керування транспортом, з розширенням для забезпечення підтримки транспортного шару СПП, таких як можливість встановлення наскрізного з'єднання IP і керування QoS, і т. д.

Приклад відповідних видів відповідальності функції керування обслуговуванням і функції керування ресурсами наведено далі. Надання заданого обслуговування кінцевим користувачам приведе до наступних дій:

– створення в функції керування обслуговуванням нового екземпляру послуги, який зв'яже результати розподілу необхідних ресурсів обслуговування і транспортних ресурсів для цього екземпляру послуги за допомогою функції керування ресурсами;

– взаємодій з функцією керування транспортним ресурсом:

- для перевірки наявності необхідних ресурсів транспортної мережі;
- для наскрізної/загальної для всіх прикладних програмних компонентів конфігурації необхідних ресурсів транспортної мережі;

- для конфігурації лінії доступу цього кінцевого користувача відповідно до технічних вимог, відповідними договору про обслуговування;
- взаємодій з функцією керування ресурсом обслуговування:
- для створення всіх даних користувача і відповідної мережевої бази даних в разі нового користувача;
- для створення у відповідній мережевій базі даних всіх відповідних даних обслуговування для цього користувача;
- для розподілу необхідних ресурсів мережі обслуговування;
- для спрацьовування/перевірки конфігурації обладнання в приміщенні користувача (CPE).

Керування ресурсом обслуговування. Функція керування ресурсом обслуговування відповідає за керування ресурсами в шарі послуг СПП.

Функція керування ресурсом обслуговування підрозділяється на функції керування мережею обслуговування і функції керування елементами обслуговування. Інфраструктура шару послуг СПП включає в себе дані, необхідні для забезпечення функціонування послуг СПП з:

- пов'язаними механізмами, що використовуються послугами для доступу до даних;
- керуванням даних.

Функція керування ресурсом обслуговування включає такі функції:

- відображення вимог функції керування обслуговуванням в профілі обслуговування і дані, що інтерпретуються ресурсами нижче;
- керування прикладним програмним забезпеченням і даними прикладного програмного компонента в мережі, включаючи представлення, оновлення, інвентаризацію, поширення, прикладні технології, відкриті інтерфейси програми та пов'язані з ними механізми забезпечення безпеки.

1.3 Мережа LTE: формальний опис структури, основні підсистеми та інтерфейси

1.3.1 Протоколи доступу LTE E-UTRAN

Швидкість передачі по LTE в низхідному напрямку (до користувача) досягає 100 Мбіт/с, у висхідному – 50 Мбіт/с. Затримка на рівні користувача не перевищує 5 мс за рахунок високої ефективності використання спектра. Такі високі характеристики забезпечуються за рахунок використання декількох антен (принцип MIMO) і мультиплексування з ортогональним поділом частот OFDM на фізичному рівні.

Мережа E-UTRAN – це найперший вузол у вдосконаленій пакетній системі EPS. Вона забезпечує високу швидкість передачі даних, малу затримку на обох площинах керування і користувача, безшовне перемикання і більше покриття комірки. Розглянемо задачі, функції та процедури рівня доступу в стеку протоколів радіодоступу.

Структура E-UTRAN показана на Рис. 1.9. Мережа складається з вузлів eNodeB (eNB), які забезпечують протоколи площини користувача (PDCP/RLC/MAC/PHY) і керування (RRC). Вузли eNB взаємодіють між собою через інтерфейс X2. Для зв'язку з удосконаленим пакетним ядром (EPC) використовується протокол S1. Обмін з вузлом керування мобільністю (MME) відбувається по інтерфейсу S1-MME, а з обслуговуючим

шлюзом (SGW) – по інтерфейсу S1-U. Інтерфейс S1 підтримує зв'язки типу безліч-безліч між MME, SGW і eNB.

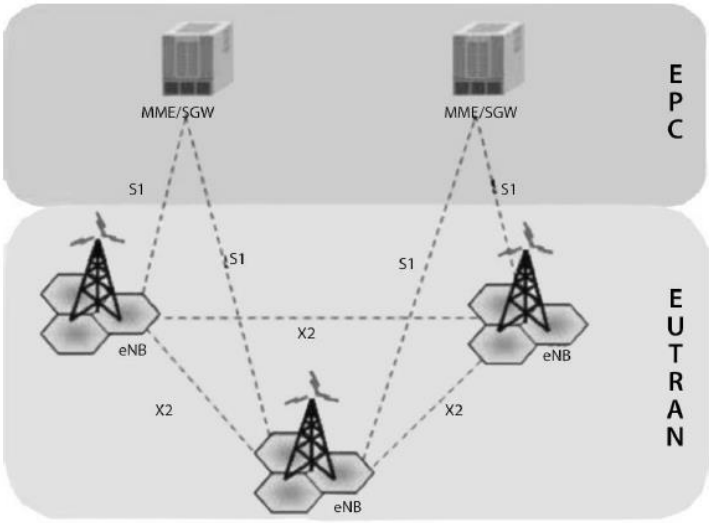


Рис. 1.9 Архітектура мережі E-UTRAN

Інформацію, яку пропускає через себе UTRAN, прийнято розділяти на два шари. До шару доступу (AS) відносяться дані, необхідні для взаємодії терміналу користувача (UE) і мережі UTRAN. Шар без доступу (NAS) містить інформацію, що переноситься між базової мережею оператора (CN) і UE через UTRAN.

Рівень доступу об'єднує протоколи радіодоступу, що забезпечують спільне використання радіоресурсів обладнання користувача і мережі доступу. Крім того, AS відповідає за з'єднання з каналом радіодоступу (RAB), за допомогою яких забезпечується взаємодія між UE і CN (сервіс NAS).

Рівень доступу надає обладнанню користувача можливість отримання доступу до ресурсів і сервісів мережі, а також всю необхідну інфраструктуру. Протоколи радіодоступу виконують такі функції:

- керування ресурсами радіоканалу (RRM). Це керування радіоканалом і радіоприйомом, контроль мобільності з'єднання і динамічний розподіл ресурсів обладнання користувача в обох напрямках передачі.
- керування трафіком:
- передача даних, в т.ч. в режимі реального часу, між інфраструктурою (рівень NAS) і обладнанням користувача;
- обробка всіх типів даних при різних параметрах каналу (рівень активності, пропускна здатність, затримка передачі і ймовірність появи помилкових бітів);

- ефективне перетворення атрибутів трафіку, які використовують не-LTE прикладними програмними компонентами, в атрибути каналу радіодоступу (RAB) на рівні доступу;
- стиснення IP-заголовка і шифрування потоків даних користувача;
- самостійний вибір MME на обладнанні користувача, коли мережа не надає відповідної інформації;
- передача даних з площини користувача на SGW;
- керування розташуванням: розподіл і передача пошукових повідомлень;
- розподіл і передача широкомовної інформації;
- задання конфігурації вимірюваних параметрів і форми виведення результатів для розподілу ресурсів і забезпечення мобільності;
- розподіл і передача повідомлень про землетруси і цунамі;
- надання первинного доступу до мережі, реєстрація та приєднання до мережі або вихід з неї;
- керування передачею на різних рівнях: між eNodeB, всередині eNodeB, між eNodeB зі зміною MME, між eNodeB зі збереженням MME, але зміною SGW, між RAT;
- функціональна різноманітність і шифрування;
- кодування радіоканалу.

SAE – це архітектура ядра мережі, розроблена консорціумом 3GPP для стандарту бездротового зв'язку LTE (Рис. 1.10).

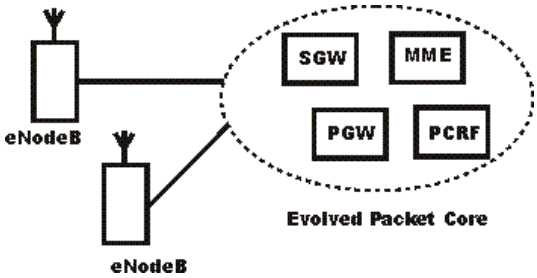


Рис. 1.10 Архітектура ядра мережі SAE

SAE є еволюційним продовженням ядра мережі GPRS, з деякими відмінностями:

- спрощена архітектура – архітектура SAE знижує експлуатаційні та капітальні витрати. Нова плоска модель означає, що потрібно підвищити пропускну здатність вузлів тільки двох типів (базових станцій і шлюзів), щоб вони впоралися з трафіком в разі його значного зростання.
- цілком побудована на IP (AIPN) – перші концепції 3G було розроблено для того, щоб голос, як і раніше передавався по системі з комутацією каналів. З тих

пір спостерігався перехід до IP-мереж. Відповідно архітектура SAE побудована на базі IP-мережі.

- забезпечує більшу пропускну здатність мережі радіодоступу (RAN) – передбачається, що низхідний канал (Down Link) буде мати швидкість понад 100 Мбіт/с, і основна увага системи буде зосереджена на мобільності смуги пропускання, мережа повинна буде підтримувати набагато більше рівнів даних.
- забезпечує меншу затримку RAN – зі збільшенням необхідних рівнів взаємодії і більш швидких відповідей, SAE концепція забезпечить рівень затримки в районі 10 мс.
- підтримує мобільність між декількома гетерогенними RAN, що включає підтримку, як систем типу GPRS, так і не-3GPP систем (наприклад WiMAX). [1]

Основним компонентом архітектури SAE є Evolved Packet Core (EPC). EPC служить еквівалентом мережі GPRS. Архітектура EPC показана на Рис. 1.11.

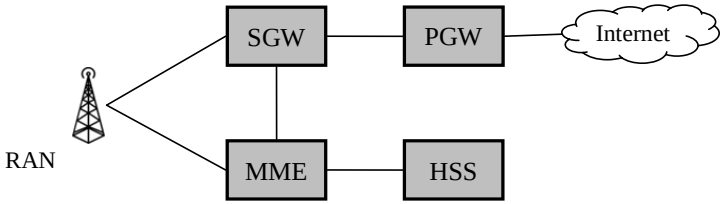


Рис. 1.11 Типова архітектура EPC/LTE

Архітектура EPC є мережею All-IP з комутацією пакетів, яка вміщує різні схеми бездротового доступу, такі як Long Term Evolution (LTE). Мережа EPC складається з різних вузлів: Evolved NodeB (eNB) вузла для радіодоступу LTE, об'єкта керування мобільністю (MME) – вузол для керування мобільності терміналу, вузла домашньої абонентської системи (HSS), вузла бази даних інформації про користувачів, вузла облслуговуючого шлюза (SGW), вузла якірної точки мобільності для керування терміналами, а також вузла шлюза пакетної мережі передачі даних (PGW), який служить шлюзом між терміналами і зовнішніми мережами, такими як Інтернет. [2]

MME – це ключовий контролюючий модуль для мережі доступу LTE. Він відповідає за процедури забезпечення мобільності, хендовера, стеження і пейджінга UE. Він бере участь в процесах активації/дезактивації мережевих ресурсів і так само відповідає за вибір SGW для UE при початковому підключенні і при хендовері всередині LTE зі зміною вузла ядра мережі (CN). Він відповідає за аутентифікацію користувача (при взаємодії з HSS).

SGW – призначений для обробки і маршрутизації пакетних даних, що надходять в підсистему базових станцій. SGW маршрутизує і направляє пакети з одними даними, в той же час виконуючи роль вузла керування мобільністю для даних користувача при хендовері між базовими станціями (eNodeB), а також вузла керування мобільністю між мережею LTE і мережами з іншими технологіями 3GPP. Коли UE вільний і не зайнятий викликом, SGW відключає низхідний канал даних (DL) і створює пейджинг, якщо

потрібно передати дані по DL в UE напрямку. Він керує і зберігає стани UE (наприклад вимоги до пропускної здатності для IP-сервісів, внутрішню інформацію мережевої маршрутизації). Він також надає копію даних користувача при узаконеному перехопленні.

PGW – забезпечує з'єднання від UE до зовнішніх пакетних мереж даних, будучи точкою входу і виходу трафіку для UE. UE може мати одночасно з'єднання з більш ніж одним PGW для підключення до декількох мереж. PGW виконує функції захисту, фільтрації пакетів для кожного користувача, підтримку білінгу, узаконеного перехоплення і сортування пакетів. Інша важлива роль PGW - бути вузлом керування мобільністю між 3GPP і не-3GPP технологіями, такими як WiMAX і 3GPP2 (CDMA 1X і EVDO).

HSS – це центральна база даних, яка містить інформацію про користувача. Функції HSS включають мобільність, керування викликами і підтримкою встановлення сеансу, аутентифікацію і авторизацію користувачів. HSS базується на pre-Rel-4 Home Location Register (HLR) і центрі аутентифікації (AuC). [1]

1.3.2 Інтерфейси між вузловими елементами в мережах стандарту LTE

Структура мережі стандарту LTE зазнала значних змін в порівнянні з мережами попередніх поколінь. Це вплинуло також і на зміну інтерфейсів між вузлами мережі. На Рис. 1.12 нижче представлена загальна модель мережі стандарту LTE та її основні інтерфейси.

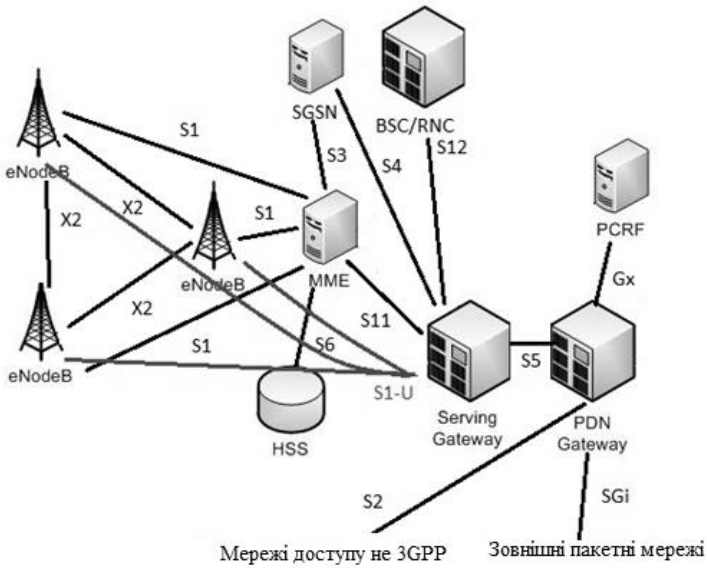


Рис. 1.12 Інтерфейси мережі стандарту LTE

Таблиця 1.1

Відображення функцій в логічну архітектуру

Function	**A/Gb mode - MS**	**Iu mode MS**	**A/Gb mode RAN**	**Iu mode RAN**	**A/Gb mode SGSN**	**Iu mode SGSN**	**Serving GW**	**GGSN**	**P-GW**	**HLR**
Network Access Control										
Реєстрація										X
Аутентифікація і авторизація	X	X			X	X				X
Контроль доступу	X	X	X	X	X	X	X	X	X	
Скрінінг повідомлень								X	X	
Адаптація до пакетного терміналу	X	X								
Збір даних					X	X	X	X	X	
Оператор, що визначає заборону					X	X				X
Пакетна маршрутизація і передача										
Relay	X	X	X	X	X	X	X	X	X	
Маршрутизація	X	X	X	X	X	X	X	X	X	
Переклад і зіставлення адрес	X	X		X	X	X	X	X	X	
Інкапсуляція	X	X		X	X	X	X	X	X	
Тунелювання				X	X	X	X	X	X	
Компресія	X	X		X	X					
Шифрування	X	X		X	X					X
Керування мобільністю	X	X			X	X	X	X	X	X
Логічне керування посиланнями										
Створення логічного посилання	X				X					
Обслуговування логічних посилань	X				X					

Керування радіоресурсом	X	X	X	X	X					

X2 – інтерфейс між eNodeB. Базові станції в мережі LTE з'єднані за принципом «кожен з кожним».

S1 – інтерфейс зв'язує підсистему базових станцій E-UTRAN і MME. По даному інтерфейсу передаються дані керування.

S1-U – інтерфейс між E-UTRAN і SAE, по якому передаються дані користувача.

S2 – інтерфейс для організації з'єднання між PDN-Gateway і мережами доступу, які не розроблялись 3GPP.

S3 – інтерфейс, що надає пряме з'єднання SGSN і MME. Він служить для передачі даних керування для забезпечення мобільності між LTE і 2G/3G мережами.

S4 – інтерфейс, що зв'язує SAE і SGSN. Він служить для передачі даних користувача для забезпечення мобільності між LTE і 2G/3G мережами.

S5 – інтерфейс між SAE і PDN-Gateway. S5 призначений для передачі даних користувача між SAE і PDN-Gateway.

S6 – інтерфейс між MME і HSS. Він використовується для передачі даних абонентського профілю, а також здійснення процедур аутентифікації в мережі LTE.

Gx – інтерфейс між PDN-Gateway і PCRF. Gx призначений для передачі правил тарифікації від PCRF до PDN-Gateway.

SGi – інтерфейс між PDN-Gateway і зовнішніми IP-мережами.

Відповідно до рекомендації 3GPP TS 23.6060 V15/0.0 (2017-09) функції рохподілені наступним чином між підсистемами забезпечення мережі LTE

1.4 Особливості віртуалізації підсистем LTE

1.4.1 Гетерогенне телекомунікаційне середовище

Гетерогенне телекомунікаційне середовище – організаційно-технічна сукупність, що складається з каналів зв’язку (фізичних та віртуальних), вузлів передачі даних (керованих апаратно та/або програмно) та обчислювального середовища, розташованого у дата центрах (як власних так і орендованих) з віртуалізованою організацією процесів обробки інформації.

Віртуалізація мережевих функцій (NFV) направлена на перетворення мережевої архітектури операторів за рахунок впровадження стандартної технології віртуалізації для об’єднання багатьох типів мережевого обладнання у високорівневі сервери, комутатори та сховища даних, які можуть бути розташовані в різних точках присутності (NFVI-PoP), таких як дата-центри, мережеві вузли та приміщення кінцевого користувача (ETSI GS NFV 001 V1.1.1(2013-10)).

Загалом, усі мережеві функції та вузли можуть розглядатись для віртуалізації і повинні бути реалізовані за стандартами. Нижче описано основні високорівневі цілі NFV:

- Швидке інноваційне обслуговування через розгортання програмного забезпечення та встановлення чіткого взаємозв’язку між мережевими функціями і кінцевими послугами.
- Покращення операційної ефективності за рахунок спільної автоматизації та робочих процедур.

- Зниження енергоспоживання за рахунок міграції робочих навантажень та вимкнення обладнання, що не використовується.
- Стандартизовані та відкриті інтерфейси між мережевими функціями та їх об'єктами керування, щоб роз'єднані елементи мережі могли забезпечуватись різними гравцями.
- Більша гнучкість у наданні віртуальних мережевих функцій для апаратного забезпечення.
- Підвищення загальної ефективності обслуговування в порівнянні з виділеними апаратними реалізаціями.

1.4.2 Віртуалізація маршрутизатора на границі локальної мережі мобільно зв'язку CPE (vE-CPE) – функціональний опис

Рішення vE-CPE покращує мережу підприємства шляхом заміни пристроїв віртуалізованими рішеннями, сумісними з NFV, розташованими як у корпоративній хмарі, так і в NFV інфраструктурі оператора. Послуги, що надаються vE-CPE, можуть включати в себе маршрутизатор, що забезпечує QoS та інші високотехнологічні послуги, такі як стабільний брандмауер L7, виявлення та попередження вторгнень тощо. Прискорювачі програм також розгортаються або як окремі пристрої, або як інтегровані сервіси маршрутизатора.

На Рис. 1.13 представлено типове телекомунікаційне підприємство, що включає штаб-квартиру з централізованою корпоративною IT-інфраструктурою та декілька філій, з'єднаних один з одним та з головним офісом підприємства. Функціональність vE-CPE може розташовуватись в різних місцях.

Рис. 1.13 Приклади розташування vE-CPE

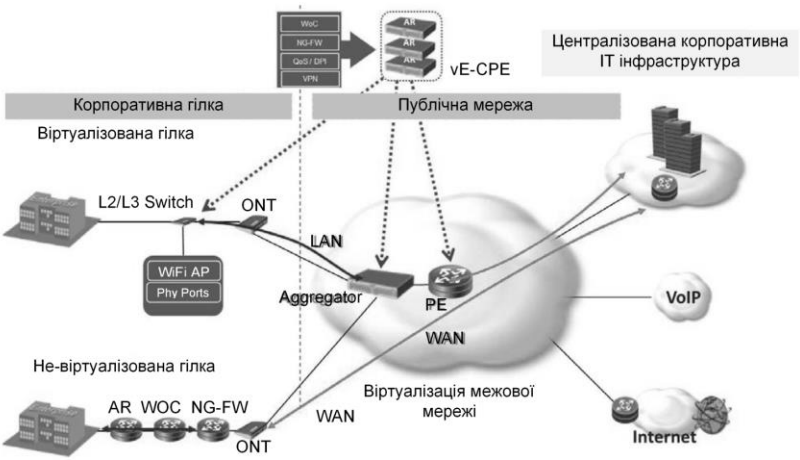


Рис. 1.14 Не віртуалізоване CPE та vE-CPE

На Рис. 1.13 представлено перерозподіл функціональних можливостей в результаті віртуалізації CPE. Локальний трафік підприємства обробляється локальним маршрутизатором L2 або L3, що забезпечує фізичний зв'язок (і, можливо, додаткову функціональність), а локальна мережа підприємства поширюється на мережу NFV оператора, розташовану у vE-CPE. Приклад функціональності, яку надає vE-CPE, включає в себе маршрутизацію, припинення VPN, підтримку QoS, DPI, NG-FW та WOC (WAN Optimizer Controller). На Рис. 1.14 порівнюється невіртуальний клієнтський сайт, який обслуговує не віртуалізований CPE, і веб-сайт, який обслуговує vE-CPE. Пунктирні фіолетові лінії вказують, де може розташовуватись функціонал vE-CPE.

1.4.3 Віртуалізація порогового маршрутизатора оператора (vPE) – функціональний опис

Віртуалізація базових маршрутизаторів може бути неможливою у короткостроковій перспективі через високі вимоги до пропускної здатності, але віртуалізація порогового маршрутизатора PE з додатковими перевагами забезпечення масштабованості послуг, наданих постачальником є більш ймовірною, за рахунок динамічної зміни об’єму або розподілу віртуальних ресурсів. Провайдер, який надає послуи віртуальної приватної мережі (PPVPNS) у vPE включає IP-VPN 3-го рівня, VPLS 2-го шару, EVPN тощо [6].

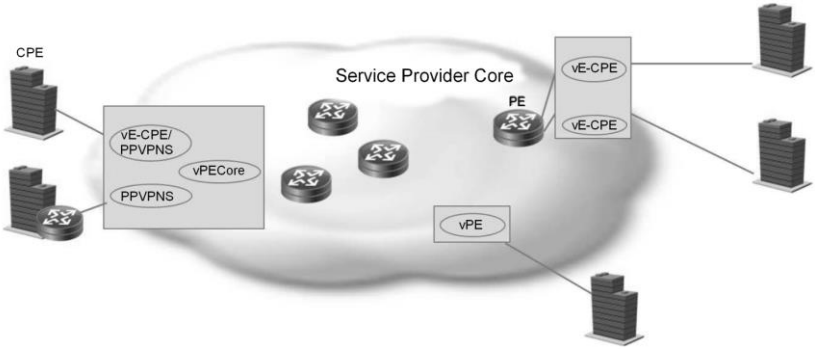


Рис. 1.15 Віртуалізація CPE, віртуальних мережевих сервісів та основних функцій порогового маршрутизатора

Віртуальні функції PE можуть:

- Інтегруватись в єдину віртуальну машину (vPE на Рис. 1.15), реалізувати еквівалентну функціональність єдиного фізичного PE або підмножину фізичних PE для масштабованості та продуктивності.
- Розбиватись на основний набір функцій (vPEcore на Рис. 1.15) та функції віртуальних мережевих служб, які включають або не включають функції CPE (PPVPNS і vE-CPE/PPVPNS).

Існує багато мережевих функцій, які зазвичай сьогодні розгортаються в корпоративних мережах в якості спеціалізованої інфраструктури обладнання, де в майбутньому постачальник послуг може надавати VNFaaS (virtual network functions as a service) для підприємства. Ці мережеві функції Enterprise включають в себе:

1) AR – маршрутизатор доступу для підприємств / Enterprise CPE
2) PE – пороговий маршрутизатор
3) FW – корпоративний брандмауер
4) NG-FW – брандмауер наступного покоління
5) WOC – контролер оптимізації мережі WAN корпоративного рівня
6) DPI – Глибока перевірка пакетів (пристрій або функція)
7) IPS – Система запобігання вторгненню та інші пристрої безпеки
8) Моніторинг продуктивності мережі.

За оцінками, велика кількість віртуалізованих пристроїв потребує підтримки з боку порогового маршрутизатора, що вимагає величезної кількості ресурсів від NFVI. VPE повинен мати можливість самостійного масштабування на площині даних та площині керування, щоб підтримувати великі таблиці переадресацій та велику кількість потоків. У порівнянні з рішеннями Virtualized Home, для віртуалізації підприємства потрібна значна менша кількість VNF, кожна з яких має набагато більшу кількість потоків та вимог до продуктивності.

Щоб обмежити вартість та масштаб, велика кількість віртуалізованих пристроїв повинна бути інтегрована на обмежену кількість процесорів.

Для досягнення цілей ефективності віртуальні функції порогового маршрутизатора повинні:

- Інтегруватись в єдину віртуальну машину;
- Розбиватись на основний набір функцій та віртуальних мережевих послуг.

VPE повинно мати можливість динамічного масштабування, щоб підтримувати велику таблицю переадресації і велику кількість потоків. Досягти масштабування vPE можна шляхом:

- Модифікації ресурсів інфраструктури, що виділяються на екземпляр vPE (наприклад, збільшення пам'яті);
- Створення додаткових екземплярів vPE.

У віртуалізованому середовищі відповідальність за належне виконання кожного сценарію лягає на постачальника VNFaaS. VE-CPE і vPE необхідні для підтримки великої кількості прикладних програмних компонентів і послуг, керованих динамікою підприємства. Крім того, під час переходу від звичайних до віртуалізованих мереж буде багато топологій та конфігурацій мереж.

Як постачальник VNFaaS, так і користувач поділяють відповідальність за керування vPE та vE-CPE. Користувачі підприємств будуть керувати та налаштовувати свої пристроїв CPE і керувати версіями SW після модернізації, навіть якщо вони є віртуалізованими та надаються в як сервіс.

VNFaaS вводить єдину точку збою, оскільки внутрішні операції з підприємством, які залежать від CPE, можуть не працювати належним чином після втрати мережевих з'єднань. Завдання полягає в тому, щоб забезпечити безперервність роботи на підприємстві під час відмови мережі або доступу до мережі (відповідно до поточної поведінки мережі).

Якщо vPE і vE-CPE повинні контролюватися централізованим контролером, слідуючи принципам і стандартам архітектури SDN, то надійне підключення між контролером та віртуальними пристроями, незалежно від їх розташування, має вирішальне значення.

Віртуалізоване середовище має гарантувати повну ізоляцію серед користувачів. Розширення корпоративної локальної мережі в операційну мережу вимагає встановлення з'єднання VPN між підприємством та віртуальною функцією оператора. Особливі заходи необхідні для захисту корпоративних даних та файлів конфігурації.

Надання VNFaaS в якості служби оцінки вимагає вимірювання показників та інфраструктури, що відповідає типу VNF, а також відповідних угод про рівень обслуговування. Значення показників якості обслуговування, отриманні за допомогою служби VNFaaS потребуватимуть відповідного аудиторського методу обліку, який буде використовуватися в якості основи для платіжних послуг.

1.4.4 Граф переадресації VNF

Граф переадресації мережевої функції (NF) [1] визначає послідовність мережевих функцій, за якою перетинаються пакети. Проста мережева служба [1] може бути реалізована в середовищі NFV за допомогою зв'язків point-to-point. Цей сценарій використання показує, що можуть знадобитись більш складні структури, такі як граф переадресації VNF (VNF FG) [1].

Графи VNF є аналогами з'єднань існуючих фізичних пристроїв кабелями. Кабелі є двонаправленими, тому більшість технологій мережевих даних (наприклад, Ethernet) будуть найближчим часом використовуватись у віртуалізованих розгортаннях. Іншими

словами, VNF Forwarding Graph забезпечує логічне підключення між віртуальними пристроями (наприклад, VNFs).

Щоб реалізувати цілі NFV, постачальникам послуг необхідно розвинути мережеві служби на абстрактному рівні, а потім розгорнути їх в екземплярах, пов'язаних з певними ресурсами NFVI (обчислювальні вузли, кінцеві точки мережі інфраструктури, існуючі фізичні NE, і т. д.).

Віртуальні мережеві функції у VNF FG мають стандартизовані інтерфейси (наприклад, L1, L2, L3, L4 та / або L7). У деяких із них, пакети мають певне призначення (наприклад, серверні функції), а в інших – ні (наприклад, Інтернет). Багато інших сценаріїв використання мають ті ж характеристики, що і сценарій використання VNF FG, а вимоги, архітектура та специфікації цих загальних характеристик повинні відповідати цілям NFV для забезпечення переходу від існуючих фізичних мережевих функцій до віртуальних аналогів.

Актори та ролі: VNF Provider – це постачальник програмного забезпечення для керування VNF та NFV; оркестрування – набір операційних систем, що підтримують NFVI.

Сценарій використання графа переадресацій VNF має наступні логічні частини та відносини між суб'єктами (Рис. 1.16):

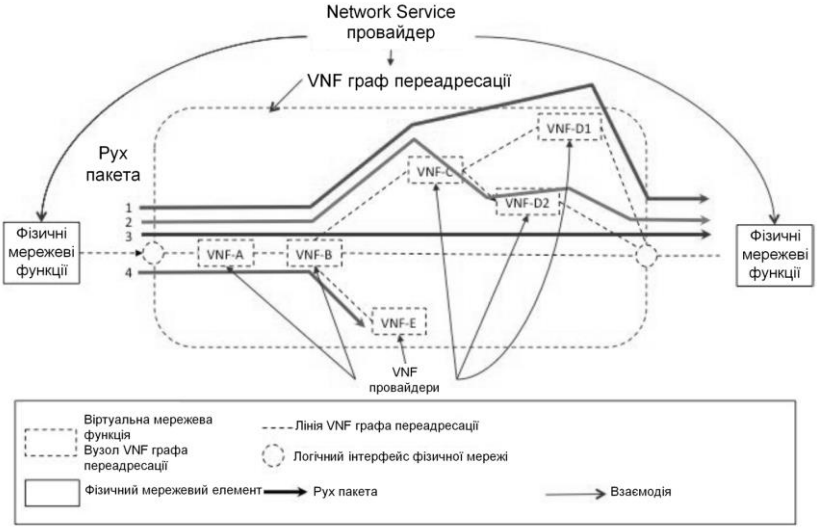


Рис. 1.16 Логіка графа переадресації VNF

Функція фізичної мережі: реалізація, яка є частиною загальної невіртуалізованої служби, що розгортається та керується постачальником послуг мережі. Це може бути фізичний доступ або магістральна мережа, автономна VM, точка з'єднання між кількома FGF VNF, що надаються різними доменами адміністратора (наприклад, NSPs).

Фізичний мережевий логічний інтерфейс: межа між VNF FG та фізичними мережевими функціями визначається постачальником послуг. Він може грунтуватися на полях у заголовку пакетів, які є джерелом або призначенням надходження пакетів від функції фізичної мережі.

Пакетний потік: вихід мережі, який є частиною загальної послуги і полягає в тому, що певні групи пакетів ідуть по тому ж шляху через VNF FG. Функціональність VNF, конфігурація та стан визначають потік пакетів через граф переадресації VNF, а пропущені VNF можуть відрізнятися в кожному напрямку для пакетів одного і того ж двонаправленого потоку.

Інфраструктура мережі NFV: забезпечує послуги з'єднання між VNF, які реалізують зв'язки графів пересилання між вузлами VNF в апаратному та/або програмному забезпеченні, як показано червоними стрілками. Вона може містити різні функції, включаючи класифікацію трафіку, інкапсуляцію тунелю, керування рухом та/або деякі форми балансування навантаження.

На Рис. 1.16 наведено приклад графа переадресації VNF, який постачальник послуг може використовувати при організації обслуговування.

У цьому прикладі постачальник послуг розробив кінцеву мережевеву службу між двома функціями фізичної мережі, що включає кілька віртуальних мережевих функцій (VNF-A, VNF-B, VNF-D1, VNF-D2, VNF-E). Ці VNF були надані одним або більше провайдерами VNF. У цих VNF є деякі метадані, пов'язані з ними, які описують основні характеристики VNF. Фактична мережева послуга являє собою набір усіх можливих пакетних потоків, які перетинають VNF FG та будь-які PNF, наприклад, як показано на Рис. 1.16. Служба мережі включає в себе інформацію (а також логіку у самих VNF), яку використовують VNF FG.

Логічний сценарій використання VNF FG пов'язаний з фізичними елементами та додатковими зв'язками між собою, як показано на прикладі Рис. 1.16. В зв’язку з цим вводиться наступна термінологія:

Фізичний зв’язок мережі: зв'язок між мережевою інфраструктурою NFV та фізичним мережевим портом мережевої функції, на кордонах між VNF та фізичними елементами.

Фізичний мережевий порт: фізичний порт на фізичній мережі або фізичний мережний комутатор/маршрутизатор або фізичний NIC.

Мережевий шлях переадресації: послідовність портів комутації обладнання та програмного забезпечення і операцій в інфраструктурі мережі NFV, який налаштовується керуванням та оркестуванням, що реалізує логічні інтерфейси «вузла» і зв'язує вузол VNF логічного графа переадресації VNF, (наприклад, VNIC на VM). Інформація VNF FG описує характеристики цих «зв'язків». Традиційні способи впровадження графів переадресації в мережі включають в себе: пересилання між фізичними пристроями на основі фізичного інтерфейсу, мережеві мости на базі VLAN, підмережі IP, конфігурації тунелів, маршрутизацію на основі політики та конкретні конфігурації BGP. Перемикач керування SDN (наприклад, OpenFlow) може реалізовувати ці традиційні методи, але

також може безпосередньо створювати графи переадресації мережі різними динамічними способами.

Середовище віртуальних машин: характеристики обчислювального та мережевого середовища для певного набору елементів програмного забезпечення VNF, налаштованих за допомогою керування та оркестування. Воно визначається інформацією, наданою постачальником VNF, та інформацією, наданою постачальником послуг мережі для VNF FG.

Постачальник послуг повинен мати можливість інвентувати всі ці VNF у своїх NFVI і прогнозувати діапазон очікуваної поведінки та продуктивності кінцевої служби мережі. На Рис. 1.17 показаний мережевий сервіс, що складається з VNF між двома фізичними мережевими функціями, при яких трафік пересилається через два фізичні пристрої та дві VNF (VNF-A, VNF-B). У цьому прикладі VNF-A є цілком віртуалізованою мережевою функцією, так як мережеве підключення також віртуалізується комутатором, однак VNF-B лише частково віртуалізується з трафіком площини даних, що проходить через фізичний перемикач.

VNF FG певного типу, де вузли та посилання мають подібну топологію з атрибутами, що визначаються параметром (наприклад, ємність, обмеження продуктивності), повинні використовувати загальний шаблон.

Граф VNF означає, що конкретний об’єкт VNF FG відповідно до цього шаблону повинен бути представлений в якості екземпляру графа переадресації NFV для набору потоків (наприклад, споживачі, підприємства, користувачі бездротового доступу, що мають доступ до Gi LAN тощо).

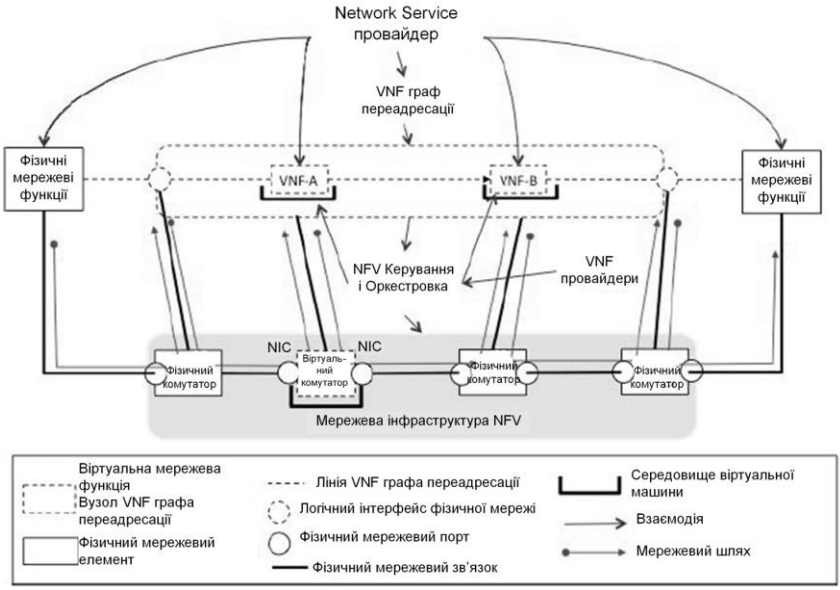


Рис. 1.17 Фізична архітектура VNF FG

Коли надається мережева служба, NFV Framework повинна вести облік використаних ресурсів інфраструктури, щоб майбутні операційні процеси (такі як локалізація несправності, відновлення, зміна розміру або припинення дії послуги) могли бути здійснені на всіх відповідних об'єктах в VNF FG.

1.4.5 Віртуалізація базової мобільної мережі та IMS

Мобільні мережі завантажені широким спектром власних пристроїв. Основна задача віртуалізації мережевих функцій полягає у зменшенні складності мережі та пов'язаних з цим проблем за рахунок стандартних технологій віртуалізації для консолідації різних типів мережевого обладнання. Така консолідація обладнання повинна зменшити сукупну вартість володінь (TCO – Total Cost of Ownership). Гнучкий розподіл мережевих функцій на такий апаратний пул ресурсів може значно підвищити ефективність роботи мережі. Це також дозволяє реагувати на збільшення попиту на певні послуги (наприклад, голосовий зв'язок), не покладаючись на механізми контролю обмеження викликів у великомасштабному сценарії стихійних лих, таких як землетрус Великого Східного Японії, під час якого мобільні мережі зіткнулися з величезною кількістю спроб дзвінків. Можливими перевагами віртуалізації мобільних базових мереж та IMS є:

- Зниження показнику TCO.
- Підвищення ефективності використання мережі завдяки гнучкому розподілу різних мережевих функцій на апаратний ресурсний пул.
- Підвищення якості обслуговування та надійності для кінцевих користувачів шляхом динамічної реконфігурації мережі, властивої технологіям віртуалізації.
- Еластичність: ємність мережі, призначена для кожної мережевої функції, може бути динамічно модифікована відповідно до фактичного навантаження в мережі, тим самим збільшуючи її масштабованість.
- Реконфігурація топології: топологію мережі можна динамічно переконфігурувати для оптимізації продуктивності.

3GPP – стандартна організація, яка визначає мережеву архітектуру та специфікації для мережевих функцій (NF) мобільних і конвергентних мереж.

В покращеному пакетному ядрі (Evolved Packet Core – EPC), що являє собою новітню мережеву архітектуру стільникової системи, прикладами мережевих функцій є MME, S/P-GW та ін.

В мультимедійній підсистемі IP (IMS), яка є архітектурою керування сеансами для підтримки надання мультимедійних послуг через EPC та інші IP-мережі, прикладами мережевих функцій є P-CSCF, S-CSCF та ін. HSS та PCRF також є мережевими функціями мережі 3GPP, необхідними для роботи в архітектурі end-to-end EPC та IMS для надання послуг. Аналогічно, онлайн-і оффлайн- системи тарифікації (OCS та OFCS) – це системи, які фіксують записи про тарифікацію як частину керування сеансами.

Цей випадок має на меті застосування віртуалізації до EPC, IMS та інших мережевих функцій, згаданих вище.

На Рис. 1.18 представлено можливий вигляд віртуалізації EPC на основі NFV.

Віртуалізовані мережеві функції (VNF), наприклад, S/P-GW, MME можуть масштабуватися незалежно від їх вимог до ресурсів. Наприклад може виникнути ситуація, коли необхідно збільшити ресурси площини користувача, не впливаючи при

цьому на площину керування, і навпаки. Крім того, VNF, що займаються площиною даних, можуть вимагати іншої кількості ресурсів NFV Infrastructure, ніж ті, що працюють тільки з сигналізацією.

Різні сценарії можуть бути увімкнені, якщо, наприклад, все ядро EPC є віртуалізованим в одному вузлі NFVI-PoP(Network Function Virtualization Infrastructure Point of Presence) або лише деякі мережеві функції є віртуалізованими.

Для того, щоб досягти необхідної безперебійної роботи служби оператора та наявності сервісу, необхідно забезпечити стійкість як на площині керування, так і на площині даних. Оскільки віртуалізація дозволяє відв'язувати NF з основного апаратного забезпечення, створення нових схем гнучкості стає можливим завдяки використанню портативних екземплярів VNF, але не обмежуючись переміщенням VM, реплікацією тощо. На Рис. 1.19 показані віртуалізовані EPC та IMS, де функції S/P-GW та IMS працюють з підключенням PDN та сеансом IMS, відповідно. Коли динамічне переміщення цих екземплярів VNF здійснюється через перевантаження або несправності віртуальної машини автоматично або за вимогою, переміщення керованих сеансів та/або підключень повинні оброблятися належним чином, щоб забезпечити безперервність роботи оператора і доступність сервісу.

Міжоператорський зв'язок та граф VNF – потенційні елементи для подальшого вивчення в даному сценарії використання.

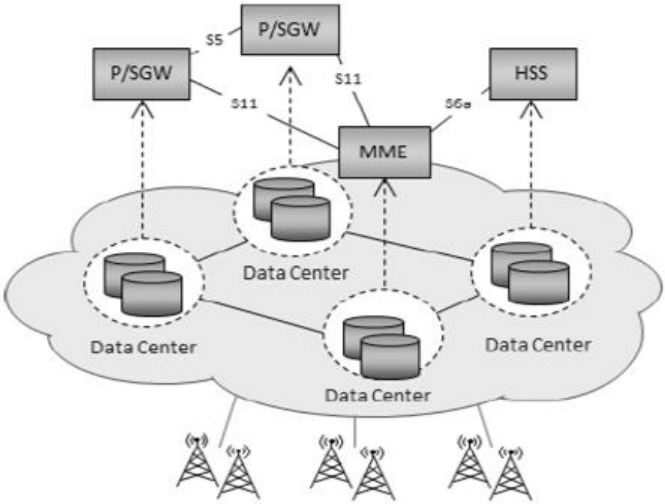


Рис. 1.18 Віртуалізація EPC

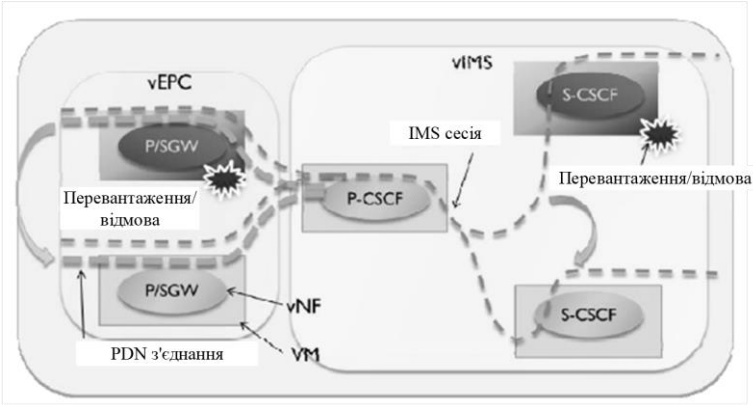


Рис. 1.19 Переміщення VNF

Віртуалізована базова мобільна мережа на основі NFV може співіснувати з невіртуалізованою (Рис. 1.20), оскільки вже розгорнуті мобільні базові мережі не функціонують на NFV. Мережеві оператори повинні мати право самостійно вибирати розгортання NFV відповідно до бажаного плану міграції з невіртуальної до віртуалізованої мережі на основі NFV.

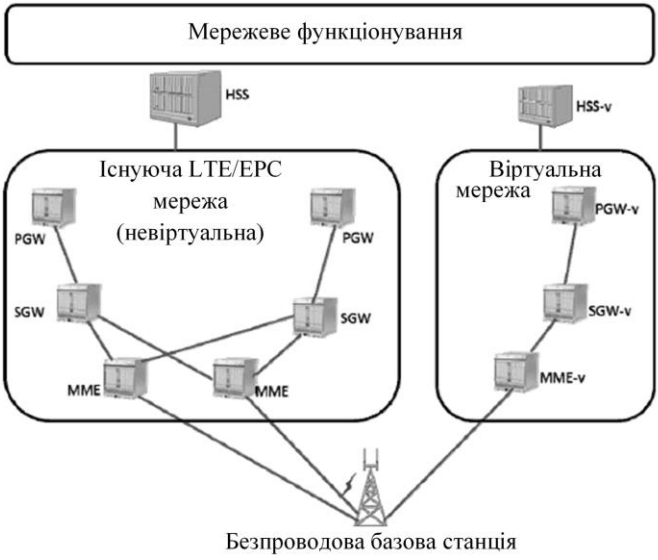


Рис. 1.20 Приклад співіснування віртуальної на невіртуальної базової мобільної мережі

Залежно від вибору оператора можуть бути різні сценарії розгортання. В якості прикладів, нижче наведено два сценарії:

- Віртуалізація деяких компонентів мобільної базової мережі. У цьому випадку віртуалізовані лише деякі мережеві функції (Рис. 1.21). Це можуть бути функції керування EPC (наприклад, MME/SGSN), HSS або сервісними вузлами (наприклад, IMS).
- Співіснування віртуалізованої та невіртуалізованої мобільної базової мережі.

У цьому випадку оператор розгортає повністю віртуалізовану базову мережу, маючи при цьому невіртуалізовану (Рис. 1.22). Віртуалізоване ядро може використовуватися для певних служб та/або пристроїв або для трафіку, що перевищує місткість невіртуалізованої мережі.

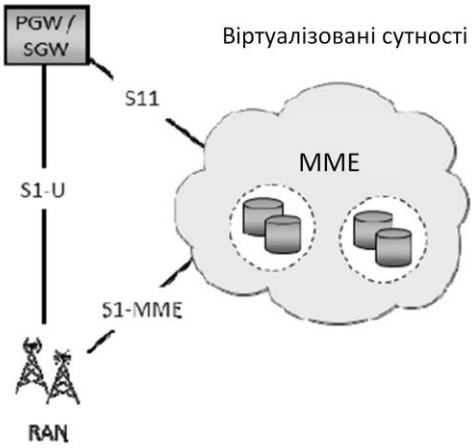


Рис. 1.21 Часткова віртуалізація базової мобільної мережі

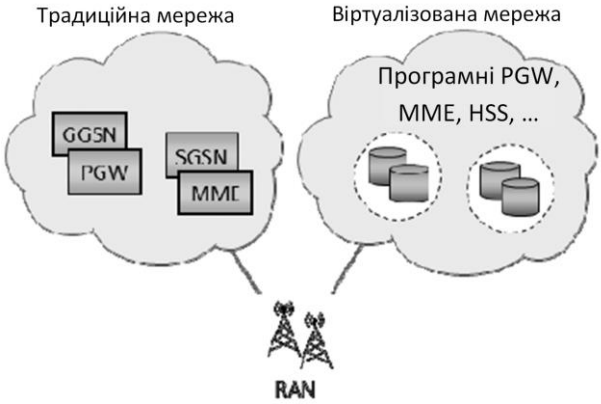


Рис. 1.22 Сервісна віртуалізація базової мережі

Для сценаріїв, що передбачають співіснування віртуалізованих та невіртуальних мобільних базових мереж потрібно визначити наступні політики:

1. Мережа радіодоступу (RAN): де збігається віртуальне і невіртуальне мобільне ядро.
2. Мережеві операційні системи: яким чином мережева операційна система для невіртуальної мережі взаємодіє з конкретними операціями базової мобільної мережі та чи потрібні нові системи підтримки операцій або вдосконалення існуючих систем.
3. Повернення до невіртуалізованої мережі: коли необхідно застосовувати механізм відмовостійкості для невіртуалізованих мережевих функцій.

Віртуалізація мобільної базової станції

Мобільний трафік мережі значно збільшується за рахунок вимог, що генеруються застосуванням мобільних пристроїв, в той час як середню кількість прибутку за одного користувача (ARPU) збільшити складно. Тому технологія 3GPP LTE направлена на задоволення потреб в підвищенні швидкості передачі даних та якості обслуговування з невисокою складністю та незмінним зниженням вартості радіодоступу і пакетного ядра. LTE також розглядається як частина доступу до радіосистем EPS (Evolved Packet System), необхідна для виконання вимог високої ефективності спектра, високої пікової швидкості передачі даних та гнучкості частоти в мережі радіодоступу (RAN). Щоб зберегти прибуток, оператори мобільного зв'язку повинні зменшувати CAPEX/OPEX, а також постійно розвивати та надавати кращі послуги своїм клієнтам. Більша частина витрат електроенергії припадає на вузли RAN. Велика кількість вузлів RAN, таких як eNodeB, зазвичай базується на власних платформах і їх недоліками є тривалий життєвий цикл розробки, розгортання та експлуатації.

Віртуалізація мобільних базових станцій використовує технологію віртуалізації IT для реалізації принаймні частини вузлів RAN на стандартних IT-серверах, сховищах та комутаторах. Це забезпечить зменшення відбитку та споживання енергії, завдяки динамічному розподілу ресурсів та балансуванню навантаження трафіку.

Крім того, NFV дозволяє створювати конкурентне середовище для постачання інноваційних сторонніх мережевих прикладних програмних компонентів, розширяючі власні межі вузлів мобільних базових станцій.

В базових мережах мобільних операторів, кілька вузлів RAN від різних постачальників, як правило, експлуатуються з різними системами мобільної мережі, наприклад 3G, LTE і WiMAX, які знаходяться в тій же зоні. Передбачається, що ці численні платформи будуть об’єднані у фізичну базову станцію (БС) на основі технологій віртуалізації.

Віртуалізація базової станції (БС) може забезпечувати обмін ресурсами між декількома логічними вузлами RAN з різних систем, динамічно розподіляючи ресурс, а також зменшуючи споживання енергії. Централізована технологія RAN (C-RAN) з віртуалізацією може сприяти більш ефективному використанню ресурсів різних фізичних БС (Рис. 1.23).

БС (використовується тут як загальний термін для позначення базових станцій 2G, 3G Node B і 4G eNode B) є частиною еталонної моделі 3GPP і включає в себе функції радіодоступу. У LTE базові станції відповідають за функції PHY, MAC, RLC (радіозв'язок), RRC та PDCP (протокол конвертації пакетних даних). Рівень PHY містить

найбільш обчислювальні інтенсивні завдання, такі як кодування/декодування каналів, FFT/iFFT.

Віртуалізація БС вимагає обробки базової радіопередачі за допомогою технологій віртуалізації, таких як високопродуктивні процесори загального призначення та віртуалізація обробки в режимі реального часу для забезпечення необхідної потужності обробки сигналу.

Крім того, віртуалізація БС для C-RAN потребує створення ресурсу обробки, тобто пулу обробки базових частот (BBU) для об'єднання ресурсів в централізоване віртуалізоване середовище, наприклад, дата центр або хмарну інфраструктуру.

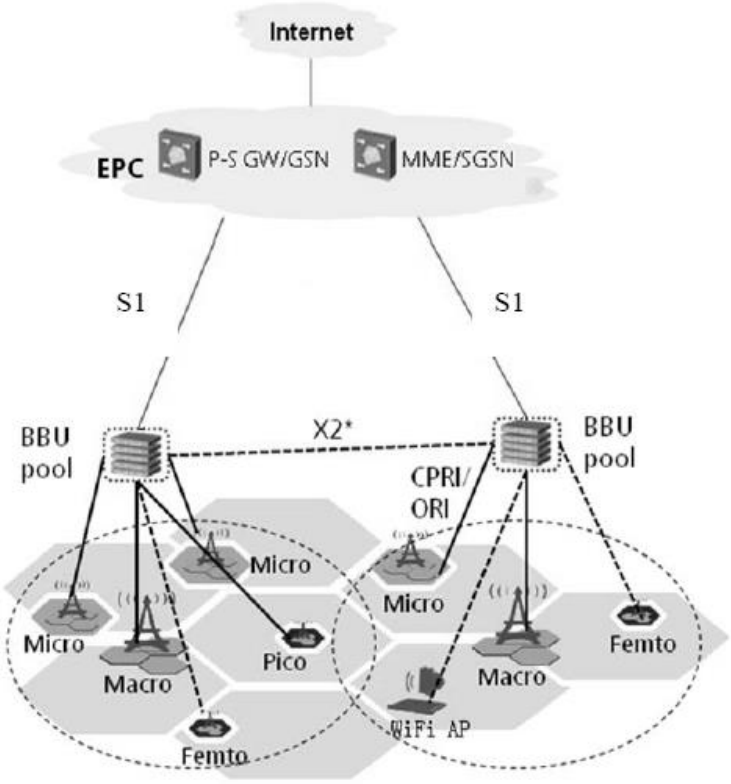


Рис. 1.23 Еволюція архітектури LTE RAN централізованим пулом BBU

Віртуалізація мобільної базової станції повинна підтримувати часткові сценарії розгортання, які враховують різні функції та елементи в частині RAN різних систем мобільної мережі. Серед можливих випадків використання виділяють:

1) Віртуалізована традиційна базова станція eNodeB та невіртуалізована eNodeB:

Віртуалізована і невіртуалізована eNodeB зв'язуються один з одним стандартним інтерфейсом X2, при цьому проблеми інтероперабельності малоймовірні, якщо вони відповідають вимогам 3GPP до затримки та джиттера.

2) Віртуалізований пул BBU та невіртуалізована eNodeB:

Віртуалізований BBU повинен мати стандартний інтерфейс X2 з невіртуальним BBU, навіть якщо інтерфейс X2 замінено власним для досягнення більш ефективного спілкування в пулі BBU. Невиртуалізована eNodeB – це фізична базова станція, географічно відокремлена від басейну BBU. У цьому випадку вищезазначені проблеми продуктивності можуть також статися через віртуалізацію та топологічну ситуацію.

Всередині вузла RAN можуть існувати цільові апаратні засоби, оскільки всі функції обробки основної смуги не можуть бути ефективно реалізовані на програмному забезпеченні. Між посилювачем та стандартною IT-платформою повинен бути реалізований інтерфейс API.

1.4.6 Віртуальна мережева функція як сервіс

Телекомунікаційні сервіси (служби) – це послуги, які надаються у гетерогенному телекомунікаційному середовищі *в процесі організації взаємодії* між абонентами або машинами.

Сьогоднішні підприємства розгортають у своїх філіалах по декілька сервісів одночасно. Для багатьох із них підтримка виділеного автономного пристрою є надто затратною та негнучкою. В інших реалізаціях, функціональність може бути забезпечена інтегрованим маршрутизатором доступу, який може бути обмеженим у наборі функцій. По мірі того, як підприємство розвивається, все більше послуг та прикладних програмних компонентів мігрують до дата-центрів або загальнодоступних хмар, що приводить до змін в побудові корпоративних мереж. Крім того, мобільність і принцип BYOD (Bring Your Own Device – принеси свій власний пристрій) стають все більш затребуваними, в результаті чого стають актуальними послуги запобігання витоку даних.

Зіткнувшись із необхідністю великих інвестицій, багато підприємств починають шукати альтернативні варіанти. Ці альтернативи можуть включати в себе віртуалізацію Enterprise CPE (маршрутизатора доступу) в мережу оператора.

Такі тенденції віртуалізації в поєднанні з перевагами, які надає NFV, забезпечують значні бізнес-можливості для постачальників послуг, які намагаються відповідати зростаючим потребам клієнтів. Традиційні IP-маршрутизатори, засновані на власному апаратному та програмному забезпеченні, є одними з найбільш капіталомістких частин інфраструктури постачальників послуг. Маршрутизатори оператора (Provider Edge routers) вичерпують ресурси платформи керування, перш ніж закінчуються ресурси платформи даних, таким чином віртуалізація функцій платформи керування покращує масштабованість.

Зберегти ресурси можна також переміщаючи функціональність маршрутизації від цільових маршрутизаторів до еквівалентних функцій, реалізованих в апаратних середовищах COTS, що забезпечують можливості хмарних обчислень, такі як NFVI.

Замість того, щоб інвестувати власний капітал у розгортання мережевої інфраструктури, постачальник послуг може надавати розширені мережеві функції. Постачальник може використовувати об’єкт VNF за допомогою інтерфейсу NFVI, який забезпечує функціональність, необхідну для впровадження клієнтського обладнання CPE та інший об’єктів VNF для платформи керування маршрутизатором оператора,

покращуючи його масштабованість. Створення функціональності VNF для підприємства в якості служби можна порівняти з поняттям хмарного обчислення – «Програмне забезпечення як сервіс».

Інститут стандартизації NIST SP 800-146 [3] визначає програмне забезпечення як сервіс (SaaS) як можливість споживачів використовувати прикладні програмні компоненти, що працюють у хмарній інфраструктурі. Споживач може керувати програмою лише з точки зору конфігурації та не може контролювати базову інфраструктуру.

У цьому прикладі віртуалізованих корпоративних служб VNF є прикладним програмним компонентом постачальника послуг. Підприємство є споживачем послуги. Підприємство не керує та не контролює NFVI або VNF. Підприємство як споживач VNFaaS не має інвестувати додатковий капітал у розширені мережеві функції, надані за допомогою платформи керування, а може отримати їх за певні кошти від постачальника послуг, якщо це необхідно. Постачальник послуг може масштабувати ресурси NFVI, виділені на екземпляр VNF у відповідь на збільшення використання VNF.

NIST SP 800-146 [3] визначили наступні переваги моделі SaaS, які також повинні застосовуватися у випадку з VNFaaS:

- незначна роль програмного забезпечення підприємства для доступу до сервісу;
- ефективне використання ліцензій на програмне забезпечення;
- централізоване керування;
- економія на попередніх витратах.

Мережі постачальників послуг Pre-NFV містять пороговий маршрутизатор PE на межі ядра перед пристроєм клієнтського обладнання (CPE), як показано на Рис. 1.24. При цьому є дві бізнес-моделі; як постачальник послуг так і підприємство можуть володіти та керувати CPE.

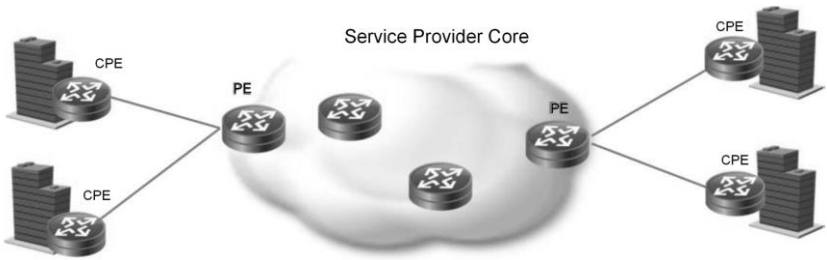


Рис. 1.24 Постачальник послуг без віртуалізації інфраструктури підприємства

Віртуалізація інфраструктури підприємства може включати:

- Віртуалізацію функцій обладнання користувача CPE (vE-CPE) у хмарі постачальника послуг.
- Віртуалізацію функцій порогового маршрутизатора оператора PE (vPE), де функції віртуальних сервісів мережі та функцій ядра PE можуть виконуватись в хмарі постачальника послуг.

Ці два етапи є незалежними і можуть розгортатись окремо. PE маршрутизатори, як правило, надаються великій кількості клієнтів, тоді як маршрутизатор CPE використовується виключно одним клієнтом. Таким чином, економія масштабу, яку можна отримати від віртуалізації CPE є значно більшою, ніж при віртуалізації PE. Отже, віртуалізація CPE є корисною як для користувачів підприємства, так і для постачальників послуг. Віртуалізація PE може відбутись на більш пізньому етапі, щоб завершити перехід до повністю віртуалізованого рішення NFV.

У деяких архітектурах vE-CPE та vPE можуть контролюватися централізованим контролером за принципами та стандартами SDN (наприклад, OpenFlow).

Постачальник послуг несе відповідальність за розгортання, налаштування, оновлення та керування роботою екземпляру VNF для забезпечення очікуваного рівня обслуговування (SLA) для абонентів VNFaaS.

1.5 Архітектура мережі LTE з віртуалізацією мережевих функцій у гетерогенній хмарній інфраструктурі

Послуги, що надаються користувачам мобільного зв'язку є різнорідними, та потребують постійного контролю показників якості обслуговування, оскільки різні типу послуг по різному чутливі до затримок та втратив процесі передачі інформаційних потоків. Саме тому для організації гнучкої телекомунікаційної системи оператору необхідно впроваджувати рішення з використанням програмного забезпечення на різних ділянках системи обслуговування. Для забезпечення контролю якості на кожному етапі необхідно контролювати та керувати не тільки процесом обслуговування, але й організації та реорганізації когнітивного ресурсу мережі доступу, що включає в себе радіо ресурс, фізичні/віртуальні обчислювальні ресурси, а також когнітивного ресурсу ядра мережі, який включає в себе фізичні та віртуальні ресурси дата центрів та телекомунікаційних мереж, що використовують для організації роботи розподілених дата центрів.

Отже, необхідно впроваджувати та вдосконалювати системи розподілу ресурсів обслуговування на всіх ділянках мережі мобільного оператору зв'язку.

За базову технологію мобільної системи зв'язку вважатимемо систему LTE, яка успішно працює вже сьогодні у провідних країнах світу, та поступово до неї наближаються українські компанії. Загальна модель мережі стандарту LTE та її основні інтерфейси показані на **Ошибка! Источник ссылки не найден.**.

Розвиток інформаційних технологій, а саме концепції хмарних обчислень, гібридні хмари, розподілені обчислення, тощо, дало поштовх до розвитку концепцій віртуалізації мережевих функцій, та програмно-керованих мереж. Мобільні мережі п'ятого покоління, які наразі розробляються науковою спільнотою усього світу, передбачають трансформацію системи зв'язку, коли функції підсистем мобільної мережі будуть частково виконуватися як прикладні програмні компоненти.

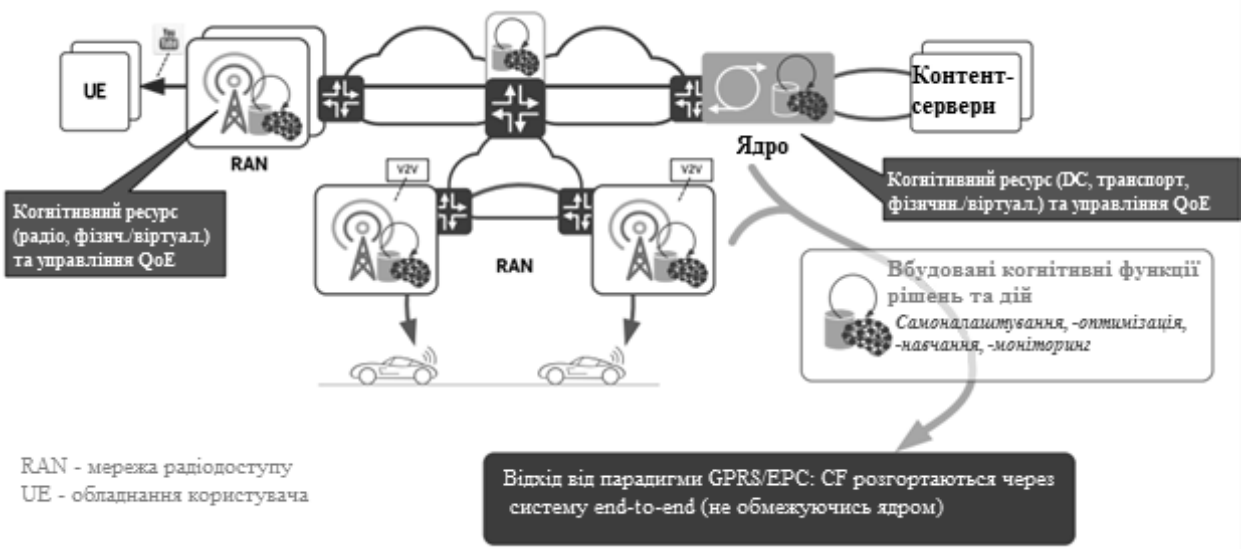


Рис. 1.25 Мережа мобільного оператора зв'язку

Архітектура мережі показана на Рис. 1.26.

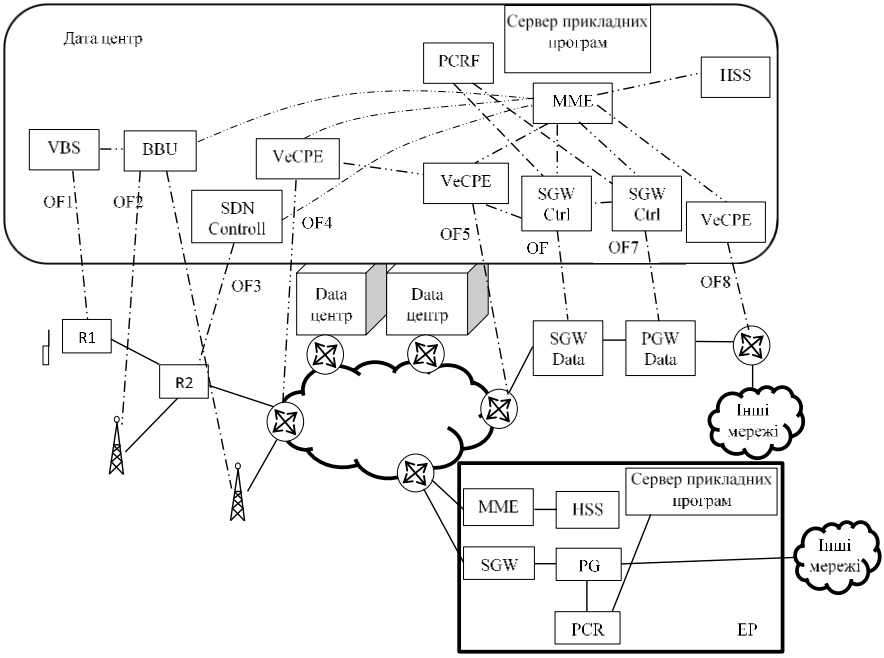


Рис. 1.26 Архітектура мережі оператора зв'язку з віртуалізацією мережевих функцій

Відповідно до специфікації ETSI GS NFV 001 v.1.1.1 (10/2013) всі обчислювальні функції, які супроводжують процес передачі будуть виконуватися у дата центрах із хмарною інфраструктурою, передбачається віртуалізація в тому числі базових станцій, передбачається створення ресурсу обробки базових частот (BBU) для об'єднання різноманітних базових станцій як віртуалізованих так і невіртуалізованих у єдине віртуалізоване середовище. Далі віртуалізація мережевих функцій відповідно до специфікації пропонується для маршрутизатора розташованого на межі локальної мережі оператора зв'язку. Маршрутизатор, що виконує функції класифікації потоків, керування маршрутизацією і забезпечення бар'єрного захисту мережі (Firewall).

Організація віртуальних базових станцій і VeCPE передбачає наявність дата центру поблизу базових станцій і кожного виходу з локальної мережі.

Таким чином, функціонування мережі оператора зв'язку являє собою розподілену територіально мережу дата центрів до кожного з яких підключені канали зв'язку, які доставляють первинну інформацію мобільних абонентів, що вимагає перетворення на найнижчому рівні (тобто сигнал, який потрібно розпізнати і розкодувати на більш високих рівнях MAC, RLC, RRC і PDCP). Таким чином можна бачити, що більшість процесів мережі оператора зв'язку відбувається в дата центрах. Мережі зв'язку стають

лише засобом доставки інформаційних повідомлень. В умовах поширення програмно-керованих маршрутизаторів маємо структуру мережі показану на Рис. 1.26.

Тут можна бачити, як мобільний абонент зв'язується з ретранслятором R1, який перетворює радіосигнал в оптичний, потім сигнал доходить до ретранслятора R2, керованого SDN контролером, який так само знаходиться в дата центрі. Потрапляючи в дата центр, сигнал обробляється віртуальною базовою станцією. Далі згідно з технологією LTE потік прямує в ядро оператора для подальшої обробки.

Підсистема BBU (блок формування модулюючих сигналів) заснована на технології програмно-конфігурованих мереж/віртуалізованих функцій мережі, це система, яка підтримує роботу не лише роботу віртуальних базових станцій, але працює для гібридних рішень 2G/3G/4G/Pre5G.

Слід зазначити, що обслуговування в ядрі визначає подальший напрямок потоку даних. Якщо потік спрямований у внутрішню мережу оператора, тоді тут же в дата центрі він направляється на обслуговування у відповідну віртуальну базову станцію, і далі направляється до абонента через ретранслятори R2 і R1. Якщо його призначення лежить за межами локальної мережі оператора зв'язку, тоді потік прямує на віртуальний маршрутизатор кордону локальної мережі, який знаходиться тут же в дата центрі. А після обслуговування потік прямує в зовнішні мережі.

Так будуть виглядати мережі наступних поколінь. Однак якщо подивитися на елемент дата центр, то стає зрозуміло, що він об'єднує групу дата центрів з'єднану через захищену мережу в єдиний логічний простір обслуговування. Забезпечення якісного обслуговування кінцевих абонентів істотно залежить від організації процесів в такому гетерогенномудата центрі, побудованому виходячи з концепції хмарних обчислень.

Згідно з рекомендацією ITU-T Y.3500 хмарні обчислення – це парадигма забезпечення мережевого доступу до масштабованого і гнучкого набору спільно використовуваних фізичних і віртуальних ресурсів з наданням та адмініструванням ресурсів на підставі самообслуговування за вимогою.

Телекомунікаційна структура описаного дата центру в якому виконується обслуговування групи функціональних блоків показаних на Рис. 1.26, представлена на Рис. 1.27, де можна бачити транспортну мережі і підключені до неї дата центри, що утворюють єдиний віртуалізований простір.

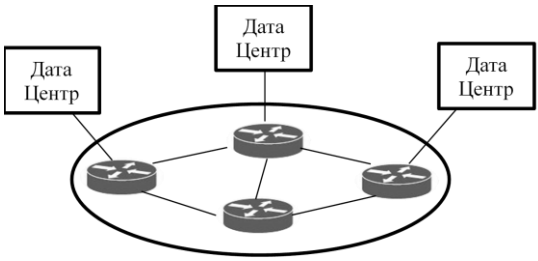


Рис. 1.27 Телекомунікаційна схема розміщення дата центрів на транспортній мережі

У мережах наведених на Рис. 1.26 ефективність обслуговування абонентів залежить від ефективності організації роботи дата центра. У даному архітектурному рішенні під дата центром розуміється складана організаційно-технічна сукупність обчислювальних та телекомунікаційних системи які забезпечують безперебійну роботу інфраструктури NFV. Ефективність роботи дата центру залежить:

- від вибору фізичних дата центів, які увійдуть до складу розподіленої структури дата центрів;
- від того, як буде вирішено питання з розміщенням мережевих функцій у інфраструктурі розподілених дата центрів;
- від організації потоків між віртуалізованими сутностями;
- від виділення ресурсів для обслуговування віртуалізованих сутностей.

Далі будуть досліджуватися особливості ефективної роботи мобільних мереж 5-го покоління, та способи забезпечення показників якості обслуговування кінцевих користувачів.

Висновки

1. Функціонування сучасного інформаційно-телекомунікаційного середовища можна охарактеризувати як роботу складної системи, що обслуговує різні типи сервісів, які надаються не лише кінцевим абонентам, але й програмно-керованим автоматизованим системам, забезпечується різнорідним обладнанням, робота якого суттєво залежить від програмного забезпечення систем керування процесом обслуговування. Системи керування процесом обслуговування сервісів набувають все більшого розвитку. Наразі для забезпечення процесу керування розробляються стандарти, моделі та підходи із застосуванням технології хмарних обчислень.
2. При трансформації телекомунікаційних систем у інформаційно-телекомунікаційні з використанням хмарних сервісів виникає необхідність у методологічному базисі організації ефективної роботи інформаційно-телекомунікаційного середовища до вимог та особливостей процесу обслуговування ТС-гібридних сервісів.
3. В розділі визначено та систематизовано підходи щодо організації процесу обслуговування у інформаційно-телекомунікаційних мережах, з частковою віртуалізацією мережевих функцій. Окреслено наступні невирішені проблеми:
    a. потрібно контролювати затримки у віртуалізованих мережевих вузлах, де інтенсивність обслуговування залежить від обчислювальних ресурсів – потребує модифікації PCRF;
    b. потрібно мати можливість реконфігурації ресурсів обслуговування – потреба у розробці нової системи керування ресурсами обслуговування гібридних телекомунікаційних сервісів у гетерогенній системі дата центрів
4. На основі проведених досліджень сформульована актуальна наукова проблема розробки методологічного базису для керування процесом обслуговування гібридних телекомунікаційних сервісів у гетерогенному телекомунікаційному середовищі, що забезпечить обслуговування кінцевих користувачів на заданому рівні якості та дозволить керувати телекомунікаційними та обчислювальними ресурсами, які забезпечують їх обслуговування.ї

5. Забезпечення якості послуг для кінцевих користувачів мереж зв'язку залежить від контролю якості на всіх етапах надання кінцевих послуг. Сьогодні, завдяки динамічно змінюваній структурі послуг, що надаються кінцевим споживачам, постійно змінюються вимоги до якості показників послуг та зростають обсяги трафіку, виникає потреба у високорозвинених системах зв'язку, які б задовольняли потреби кінцевих користувачів.

Контрольні запитання

1) Що таке NFV?
2) Що таке VNF?
3) З якими факторами та тенденціями пов’язана робота телекомунікаційних систем, які надають послуги зв’язку?
4) Що таке NGN та базова еталона модель NGN?
5) Як виглядає загальна функціональна архітектура NGN та які функції в неї входять?
6) Які рівні виділяють в узагальненій архітектурі NGN?
7) Що таке хмарні компоненти?
8) Які існують стандарти для забезпечення керування телекомунікаційними системами?
9) Які є особливості керування сервісами із застосуванням хмарних обчислень?
10) З чого складається система керування хмарними послугами?
11) Що таке гібридний сервіс?
12) Які вимоги висуваються до хмарної системи керування телекомунікаціями?
13) З чого складається функціональна платформа керування сервісами в хмарній системі керування телекомунікаціями?
14) Які існують функціональні вимоги до керування сервісами в хмарній системі керування телекомунікаціями та коротко описати їх.
15) З яких двох головних підфункцій складається функція керування ресурсами?
16) Що таке мережа E-UTRAN та яка її архітектура?
17) Що таке мережа SAE та яка її архітектура?
18) Які функції виконує EPC в архітектурі SAE?
19) Поясніть загальну модель мережі стандарту LTE та її основні інтерфейси.
20) Як розподілені функції між підсистемами мережі LTE?
21) Що таке гетерогенне телекомунікаційне середовище?
22) Які існують основні високорівневі цілі NFV?
23) Яку функціональність надає CPE маршрутизатор?
24) Де розташовується vE-CPE?
25) Поясніть віртуалізацію порогового маршрутизатора.
26) Які мережеві функції зазвичай розгортаються в корпоративних мережах?
27) Що таке граф переадресації мережевої функції?
28) Як виглядає фізична архітектура VNF FG?
29) Які переваги віртуалізації мобільних базових мереж та IMS?
30) Які можливі сценарії має віртуалізація EPC на основі NFV?
31) Які політики потрібно визначити для сценаріїв, що передбачають співіснування віртуалізованих та невіртуальних мобільних базових мереж ?

32) Що таке телекомунікаційний сервіс?

33) Які переваги моделі SaaS (програмне забезпечення як сервіс)?

34) З чого складається мережа мобільного зв’язку?

35) Які особливості архітектури мережі оператора зв’язку з віртуалізацією мережевих функцій?

36) Від чого залежить ефективність роботи дата центру?

РОЗДІЛ 2
КОНФІГУРАЦІЯ РЕСУРСІВ МЕРЕЖЕВИХ ФУНКЦІЙ В ГЕТЕРОГЕННОМУ СЕРЕДОВИЩІ

2.1 Проблематика віртуалізації мережевих ресурсів

Сьогодні мобільні абоненти бажають залишатися на зв'язку в будь-якому місці, в будь-який час, і використовуючи будь-який пристрій. Це явище спонукає операторів мобільного зв'язку до побудови складних мережевих архітектур з включенням нових можливостей і розширень, якими важче керувати [7]. Два поняття знаходяться в центрі досліджень і розробок в даний момент, а саме Віртуалізація Мережевих Функцій (Network Functions Virtualization – NFV) і Програмно Конфігуровані Мережі (Software Defined Networking – SDN) [8].

Існуючі мобільні мережеві інфраструктури складаються з виділених мережевих вузлів, які зазвичай розміщуються в різних точках мережі, і кожному вузлу призначається надання певного набору функцій і сервісів. Така архітектура є негнучкою з точки зору впровадження нових сервісів, здійснює неоптимальну маршрутизацію трафіку і неефективне використання ресурсів мережі. Крім того, масштабованість і економічність такої архітектури стає проблемою у світлі останніх прогнозів трафіку і очікувань користувачів [9].

З поширенням потужних мобільних пристроїв (наприклад, смартфонів, планшетних ПК і ноутбуків), а також зі зростаючою популярністю мобільних мультимедійних прикладних програм, очікуються підвищені вимоги до пропускної здатності. Згідно з [10] очікується, що загальний мобільний трафік даних зросте до 77 екзабайт на місяць до 2022 року, що в сім разів більше в порівнянні з 2017. Мобільний трафік даних буде рости з середнім темпом річного зростання (CAGR) у 46 відсотків з 2017 до 2022 (рис. 2.1).

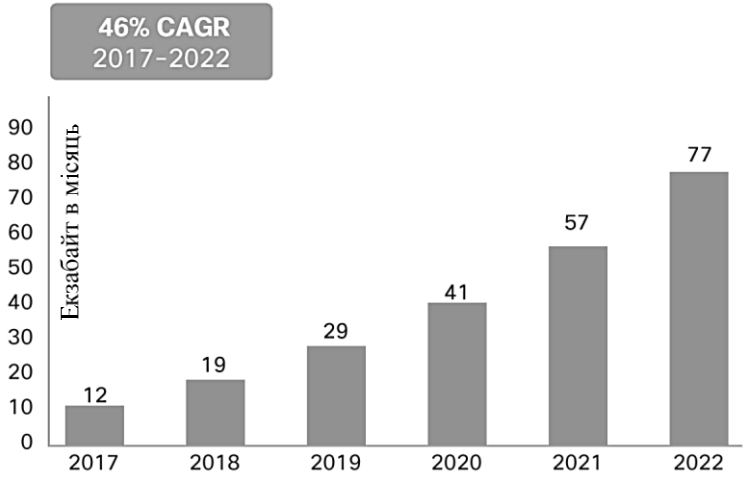


Рис. 2.1 Прогноз Cisco трафіку мобільних даних до 2022 [10]

У мережах майбутніх поколінь, зв'язок machine-to-machine (M2M) буде поширюватися на широке коло речей, таких як транспортні засоби, побутова техніка, і інші подібні об'єкти, що призводить до Інтернету речей (Internet of Things – IoT) [11], комунікацій доповненої реальності, віртуальної реальності, а також інших передових технологій. Крім того, сервіси та прикладні програми будуть більш складними і диверсифікованими [12].

Як показано на рис. 2.2, транспортні засоби, будинки, персональні пристрої, роботи, датчики, і так далі всі будуть з'єднані безпровідними мережами. Тобто, буде досягнута автоматична і інтелектуальна система керування і контролю. Зростання числа пристроїв буде впливати на ринок IoT, який, за оцінками, має розмір у 19 трильйони доларів США [13]. Число пристроїв, як очікується, досягне 50 мільярдів [14]. Крім того, повинні надаватися сервіси з багатим контентом, такі як потоковий перегляд фільмів у режимі реального часу, які потребують високої роздільної здатності, та телехірургія, що вимагають малого часу затримки [15].

<table>
<tr><td colspan="2">Все безпровідно з’єднано
Моніторинг/збір інформації & пристрої контролю</td><td colspan="2">Розширення/збагачення безпроводових сервісів
Багатий контент в реальному часі & гарантії безпеки</td></tr>
<tr><td>Множинні персональні пристрої
Взаємодія між багатьма пристроями</td><td>Транспортування (Машина/Автобус/Потяг)
Розваги, навігація, інформація трафіку</td><td>Потокове відео
4K/8K відео роздільна здатність
Відео на газетах
Фонове відео</td><td>Новий тип терміналу
Окуляри/Дотиковий Інтернет</td></tr>
<tr><td>Електроніка споживача
Віддалена робота використовуючи персональний термінал</td><td>Годинник/прикраси/одяг
Інтерфейс людини та сенсори догляду за здоров’ям</td><td>Охорона здоров’я
Віддалена перевірка здоров’я та консультація</td><td>Освіта
Віддалене навчання
Будь-який урок будь-де/в будь-який час</td></tr>
<tr><td>Будинок
Віддалений контроль обладнання
Безпека дому</td><td>Сенсори
Розумна електромережа
Фермерство
Автоматизація фабрик
Погода/середовище</td><td>Хмарні обчислення
Всі види сервісів підтримувані мобільною персональною хмарою</td><td colspan="2">Система безпеки та рятування
Попередження нещасних випадків
Стійкість до лих</td></tr>
</table>

Рис. 2.2 Сервіси в епоху мереж майбутніх поколінь[15]

На рис. 2.3 [16] області A та B являють собою поточні мобільні послуги, що базуються на стільникових мережах. З іншого боку, область C представляє майбутні послуги на базі «Інтернету речей» (IoT), де використовується величезна кількість малопотужних терміналів, тоді як послуги в області D, такі як віддалена хірургія, будуть мати суворі вимоги до продуктивності (такі як низька затримка, висока надійність і невелика кількість терміналів), і будуть надаватися через майбутні базові мережі. Ця тенденція показує, що оператори мобільного зв'язку повинні відповідати таким вимогам сторонніх сервіс-провайдерів, таких як автомобільна промисловість, при збереженні операційних витрат на розумному рівні.

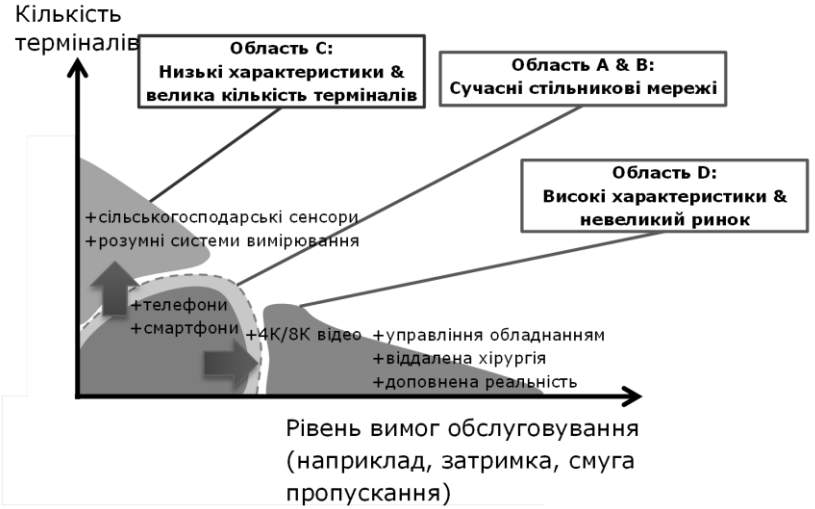


Рис. 2.3 Розширення сфери діяльності оператора [16]

Останні нововведення в мобільних телекомунікаційних технологіях та мобільних терміналах стимулюють розповсюдження різних сервісів із широким діапазоном вимог щодо затримки, мобільності та надійності серед інших [17].

Надання послуг в телекомунікаційній галузі традиційно ґрунтується на тому, що мережеві оператори впроваджують фізичні пропрієтарні пристрої та обладнання для кожної функції, яка є частиною певного сервісу. Крім того, сервісні компоненти мають чіткі ланцюги і/або порядок, які повинні бути відображені в топології мережі і в локалізації сервісних елементів. Це, в поєднанні з вимогами до високої якості, стабільності і строгим дотриманням протоколу, привело до тривалих циклів продукту, дуже низької гнучкості обслуговування і сильної залежності від спеціалізованих апаратних засобів.

Проте, вимоги користувачів до більш різноманітних і нових (короткоживучих) послуг з високими швидкостями передачі даних продовжують збільшуватися. Таким чином, телекомунікаційні сервіс-провайдери (Telecommunication Service Provider – TSP), повинні відповідним чином і постійно набувати, зберігати і експлуатувати нове фізичне обладнання. Це не тільки вимагає високих і швидко змінюваних навичок для техніків що експлуатують та управляють цим обладнанням, а й вимагає щільного розміщення мережевого обладнання, такого як базові станції. Все це призводить до високих капітальних і експлуатаційних витрат для TSP.

Більш того, навіть з цими високими вимогами абонентів, зростання капітальних і експлуатаційних витрат не може бути переведене в більш високу абонентську плату, так як TSP відомо, що через високу конкуренцію, як між собою, так і від over-the-top послуг на каналах передачі даних, підвищення цін призводить лише до відтоку клієнтів. Тому

TSP були змушені шукати шляхи побудови більш динамічних і сервіс-обізнаних мереж з метою скорочення життєвих циклів продукції, операційних і капітальних витрат і підвищення оперативності обслуговування [18].

Значна залежність мереж від апаратного забезпечення та існування різних спеціалізованих апаратних пристроїв, таких як брандмауери, обладнання глибокої інспекції пакетів (DPI), і маршрутизаторів в мережевій інфраструктурі, посилили проблеми, що стоять перед провайдерами послуг мережі [19].

Як правило, оператори мобільного зв'язку справляються з підвищеними навантаженнями трафіку шляхом розширення/покращення загальної пропускної здатності мережі відповідним чином. Проте, це все більш і більш важко реалізувати за рахунок збільшення капітальних/операційних витрат (CAPEX/OPEX) в світлі низької рентабельності інвестицій (ROI). Окрім низького ROI, надмірне резервування ресурсів більше не вважається життєздатною стратегією, щоб задовольнити збільшення трафіку, так як відповідно до [20] до 80% обчислювальної потужності базових станцій і до половини потужності ядра мережі є невикористаними. Це призводить до низького використання мережевих ресурсів, а також до високого рівня споживання енергії, що знижують економічну ефективність мережі для операторів мобільного зв'язку [9].

Принцип NFV спрямований на перетворення мережевих архітектур шляхом впровадження мережевих функцій в програмному забезпеченні, що може працювати на стандартній апаратній платформі. Крім того, він спрямований на перетворення традиційних мережевих операцій, оскільки програмне забезпечення може бути легко переміщене, або створено сутність в різних місцях без необхідності використовувати нове обладнання. NFV має багато переваг, від поліпшення операційної ефективності і зниження енергоспоживання до коротших інтервалів розгортання/оновлення і майже оптимального використання мережевих ресурсів, оскільки будівельні блоки можуть виділятись і перерозподілятись під час виконання в залежності від вимог [21, 22].

Хмарна мобільная мережа це [23]: мережа мобільного зв'язку + децентралізовані обчислення + смарт зберігання, що пропонується як один сервіс, який надається за вимогою, еластично та з оплатою Pay-As-You-Go.

Ідеєю хмарної мобільної мережі є розширення концепції хмарних обчислень за межі центрів обробки даних у напрямку до мобільного кінцевого користувача (як показано на рис. 2.4).

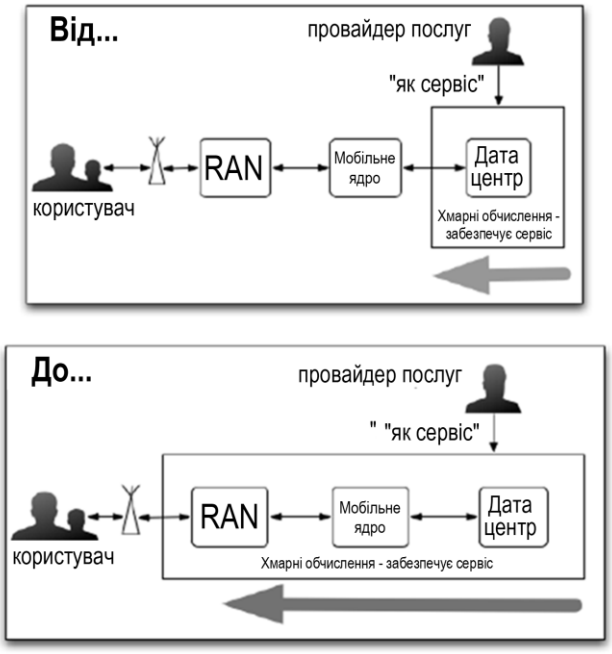


Рис. 2.4 Розширення концепції хмарних обчислень за дата центри до мобільного кінцевого користувача [23]

NFV прокладає шлях до ряду відмінностей в способах реалізації надання мережевого сервісу в порівнянні з існуючою практикою. Таким чином, ці відмінності можна охарактеризувати наступним чином [18]:

1. Відв'язка програмного забезпечення від апаратного. Оскільки елемент мережі більше не є об'єднанням інтегрованих апаратних і програмних сутностей, еволюція обох є незалежною один від одного. Це дозволяє мати окремі терміни розробки і технічного обслуговування програмного і апаратного забезпечення.
2. Гнучке розгортання мережевих функцій. Відрив програмного забезпечення від апаратного допомагає перерозподілити і спільно використовувати ресурси інфраструктури, таким чином, разом, апаратне і програмне забезпечення, може виконувати різні функції в різний час. Це допомагає мережевим операторам розгортати нові мережеві сервіси швидше по тій же фізичній платформі. Таким чином, компоненти можуть бути створені в будь-якому NFV-сумісному пристрої в мережі і їх з'єднання можуть бути встановлені на гнучкій основі.
3. Динамічне масштабування. Розділення функціональності мережевої функції на створювані програмні компоненти забезпечує більшу гнучкість масштабування реальної продуктивності віртуальної мережевої функції (Virtual Network Function – VNF)

більш динамічно і з більшою деталізацією, наприклад, відповідно до фактичного трафіку, для якого оператор мережі повинен надавати ємність.

Таким чином, метою NFV є трансформація способу, яким оператори мереж і провайдери мережевих послуг будують, керують та розгортають мережеву інфраструктуру, завдяки еволюції технологій віртуалізації. Це перетворення виконується за допомогою консолідації різних типів віртуалізованих мережевих функцій в стандартних комп'ютерах загального призначення (серверах, пристроях зберігання даних і т.д.), які можуть бути розташовані в дата центрах, мережевих вузлах і близько до приміщень кінцевого користувача (див. рис. 2.5) [24].

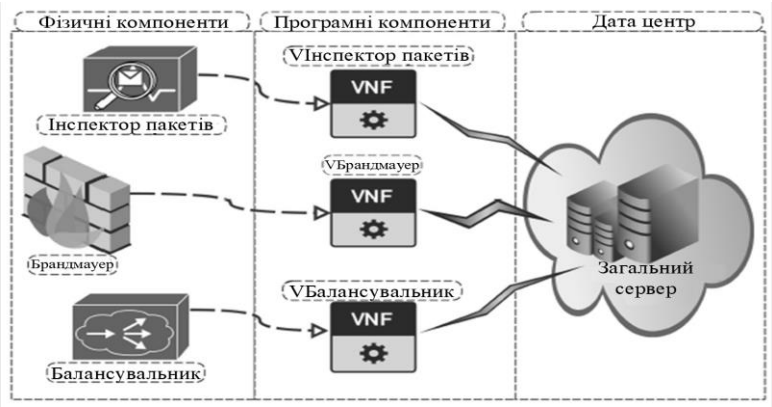


Рис. 2.5 Середовище NFV

З масовим розповсюдженням LTE оператори стикаються з сигналізацією, яка загрожує завалити традиційні базові мережі. Оскільки абоненти стають дедалі активнішими в соціальних мережах і одночасно залучають все більше прикладних програм, смартфони можуть створювати величезні обсяги сигналізації, коли вони взаємодіють із мережею.

Проникнення мобільного широкосмугового зв'язку буде збільшуватися, і як наслідок, зростання трафіку сигналізації значно перевищить відповідний приріст трафіку даних. Компанія Nokia Siemens Networks [25] прогнозує, що зростання трафіку сигнальних повідомлень буде на 50% швидше, ніж зростання трафіку даних протягом найближчих кількох років. Проте збільшення використання смартфонів є лише одним із факторів вибуху сигнального трафіку.

Спонукальні чинники сигнального трафіку (рис. 2.6):

- оператори все більше покладаються на поінформованість про обслуговування, щоб забезпечити кращий досвід для користувачів смартфонів; у поєднанні з диференційованою тарифікацією, результат у збільшенні сигналізації становить до 30% (див. рис. 2.7);

- зростаюча кількість підключених до Інтернету мобільних пристроїв міжмашинної взаємодії та прикладних програм із високими вимогами мобільності призведе до значної сигналізації;
- Voice over LTE вимагає масштабованої пропускної здатності сигналізації та диференційованої якості сприйняття сервісів в режимі реального часу;
- виникаючі тенденції та нові бізнес-моделі вимагають масштабованості площин користувача та управління;
- зростаюче використання онлайн-маркетингу, хмарного сховища, контент-бізнесу та прикладних програм;
- збільшення кількості клієнтів, які мають декілька персональних пристроїв;
- конвергенція таких галузей, як медіа, соціальні мережі, охорона здоров'я, енергетика та екологічні послуги тощо.

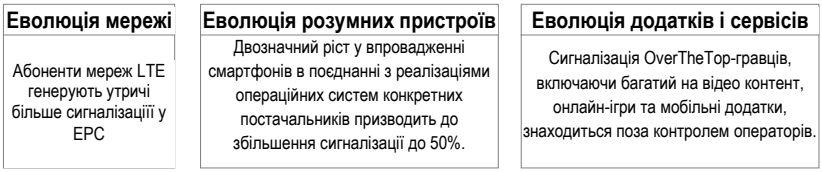


Рис. 2.6 Різноманітні причини зростання трафіку сигналізації

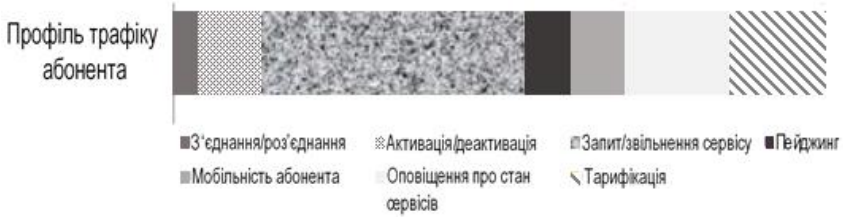


Рис. 2.7 Поінформованість щодо сервісів та диференціація тарифікації можуть додати додаткові 30% до загального навантаження сигналізації у мережах

У мережі 3G, контролер радіомережі (Radio Network Controller – RNC), знаходиться між базовою станцією і елементами базової мережі, ефективно захищаючи базову мережу від великої кількості сигналізації, що генерується мережею радіодоступу для керування мобільністю. На відміну від цього, LTE використовує пласку архітектуру, яка усуває RNC. Оскільки базова мережа підключена безпосередньо до базових станцій LTE, то вона повинна обробляти весь трафік сигналізації (див. рис. 2.8).

Середня вимога сигналізації на одного абонента до 42% вище у LTE порівнюючи з HSPA. Зрозуміло, що сигналізація стала вирішальним фактором при визначенні розмірів базової мережі [25].

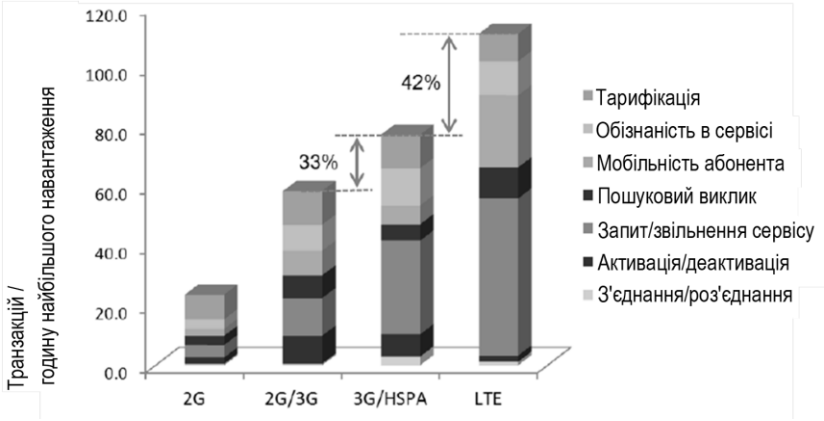


Рис. 2.8 Середні вимоги сигналізації на одного абонента [25]

Еволюційних шлях архітектури базової мобільної мережі представлено на рис. 2.9 [26].

Однією з основних переваг NFV є обіцянка підвищити ефективність використання енергії в результаті консолідації ресурсів, а також їх більш динамічного використання. У [27] визначили, що частина мобільної мережі з найвищими перспективами використання енергії є Evolved Packet Core (EPC), де віртуалізація функцій призводить до скорочення на 22% споживання енергії та підвищення на 32% в області енергоефективності (рис. 2.10).

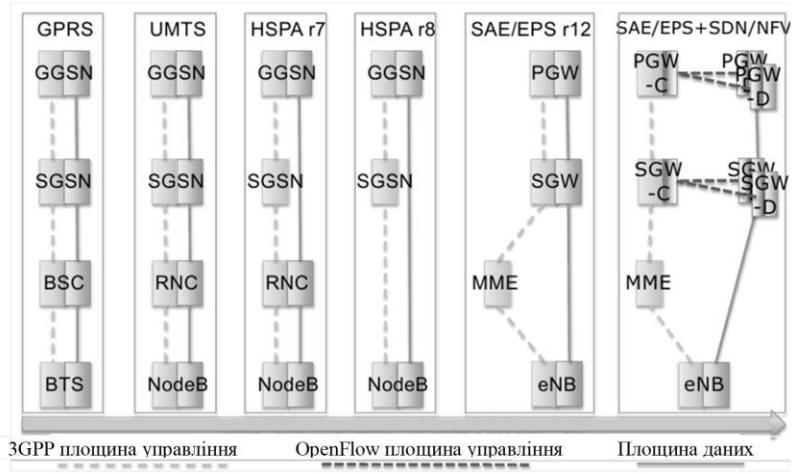


Рис. 2.9 Еволюційних шлях архітектури ядра мобільної мережі

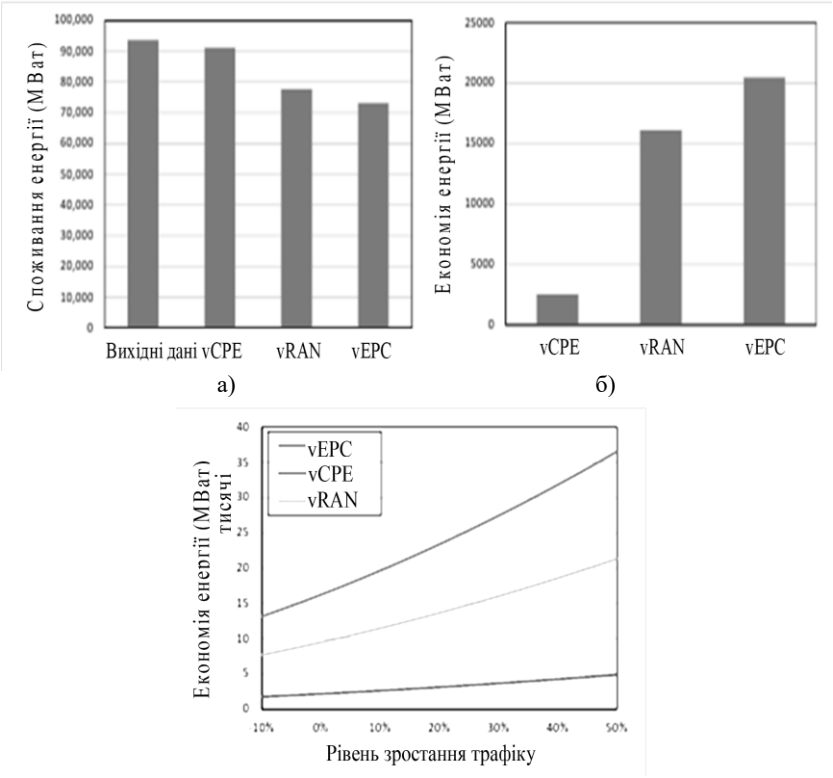


Рис. 2.10 Енергоспоживання у віртуалізованих мережах а) загальне енергоспоживання б) економія енергії в) варіація економії енергії [27]

У EPC, яка є останньою архітектурою базової мережі для системи стільникового зв'язку, до прикладів мережевих функцій відносяться MME, S/P-GW, і т.д. HSS і PCRF є іншими мережевими функціями 3GPP, які необхідні в архітектурі для надання сервісу з кінця в кінець. Аналогічним чином, онлайн і оффлайн системи тарифікації (OCS і OFCS) представляють собою системи, які фіксують записи тарифікації в рамках управління сесією.

Таким чином, одним із варіантів застосування NFV є віртуалізація EPC [28].

З часом EPC стане загальною базовою мережею для послуг з комутацією пакетів у мережах 2G та 3G, а також потенційно для Wi-Fi. EPC є центральним елементом архітектури мобільної мережі та є важливим для надання сервісів кінцевим користувачам. Вона розташована між мережею радіодоступу (RAN), IP-мережею,

зовнішніми сервісами та низкою інших мережевих елементів, що використовуються для автентифікації, захисту та надання послуг кінцевим користувачам. Це стратегічна точка контролю за наданням послуг мобільної передачі даних, як показано нижче на рис. 2.11 [29].

Рис. 2.11 EPC як важлива складова мобільної мережі [29]

EPC є однією з ключових областей, яку оператори та вендори визначили як хороший кандидат для NFV. Наприклад, визначається як життєздатний і привабливий випадок використання в специфікаціях робочої групи ETSI.

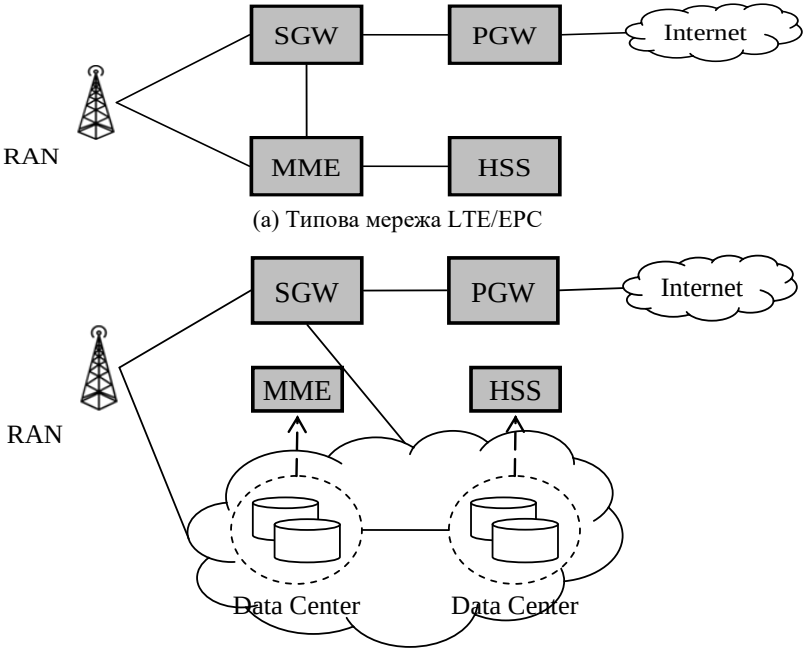
У [29] також виявлено великий інтерес до віртуальних мобільних базових мереж серед операторів через ряд власних дослідницьких проектів та залучень операторів, здійснених останнім часом. Є кілька причин, чому EPC вважається гарним кандидатом для віртуалізації. Комерційно, це нова та така, що розширюється, інвестиція, яке все ще має тривалий час життя попереду, тому узгодження інвестицій в мережу з майбутніми технологічними тенденціями є логічним. Технічно, природа EPC та порівняно скромні обсяги трафіку в мобільних мережах підходять для віртуалізації, хоча це дещо залежить від конкретної моделі застосування та розгортання.

Враховуючи важливу роль EPC у забезпеченні гарантованого надання послуг з найкращою якістю сприйняття (Quality of Experience – QoE) та оскільки мобільна мережа зазнає трансформацію, а механізми надання послуг удосконалюються, операторам мобільних мереж необхідно розуміти, визначати та вимірювати економічні вигоди від віртуалізації. Деякі з них [30]:

- показано, що віртуалізація всього комплексу функцій EPC дає змогу інфраструктурі мобільної мережі працювати на більш високому рівні використання до 87%, що безпосередньо призводить до підвищення ефективності експлуатаційних витрат до 25%;
- гнучкий рівень оркестровки може скоротити час виходу на ринок сервісу (Time To Market – TTM) шляхом спрощення критичних засобів організації, розгортання, автоматизації, взаємодії, настройки та оптимізації, а також всебічної доступності мережевих функцій EPC. Зокрема, в дослідженні [30] було виявлено

зниження TTM на 67% для запуску нових сервісів, таких як MVNO, приватна мобільна мережа, випадки використання IoT та VoLTE.

На практиці, більш дорогі спеціалізовані апаратні засоби часто працюють швидше і ефективніше ніж віртуалізовані сутності, навіть хоча останні є більш гнучкими. Оскільки спеціалізовані апаратні засоби на даний час широко використовуються, цілком імовірно, що гібридні сценарії розгортання стануть поширеними, коли частина сервісів надається фізичним обладнанням (рис. 2.12). У NFV мережах, набір ланцюгів сервісів повинен розташовуватись на фізичних вузлах мережі. Ланцюг сервісів – це набір з одного або декількох сервісів або віртуальних машин, які з'єднані разом для забезпечення певної функціональності, і можуть бути представлені у вигляді графа, що містить сервіси і мережеві вимоги між цими сервісами. У гетерогенному мережевому середовищі, ланцюги сервісів можуть розміщуватись або з використанням фізичного обладнання, або з використанням віртуалізованих сутностей. Успіх цього підходу залежить від наявності та продуктивності алгоритмів, що визначають де і як ці структурні блоки створюються [22].



(а) Типова мережа LTE/EPC

(б) Часткова віртуалізація LTE/EPC

Рис. 2.12 Мобільна базова мережа

Поточна архітектура EPC показана на рис. 2.12(а). Архітектура EPC є мережею IP з комутацією пакетів, яка складається з різних вузлів: Evolved NodeB (eNB) вузла для радіодоступу LTE, вузла управління мобільністю (Mobility Management Entity – MME) – вузол для управління мобільності терміналу, вузла домашньої абонентської системи (Home Subscriber System – HSS), вузла бази даних інформації про користувачів, вузла облслуговуючого шлюза (Serving Gateway – SGW), вузла якірної точки мобільності для управління терміналами, а також вузла шлюза пакетної мережі передачі даних (Packet Data Network Gateway – PGW), який служить шлюзом між терміналами і зовнішніми мережами, такими як Інтернет [15].

2.2 Огляд архітектури LTE/EPC

На рис. 2.13 показана типова архітектура LTE/EPC, яка показує мережеві елементи, які з'єднані між собою через добре визначені інтерфейси, і кожному елементу мережі присвоюється виділений набір спеціалізованих функцій/сервісів [9].

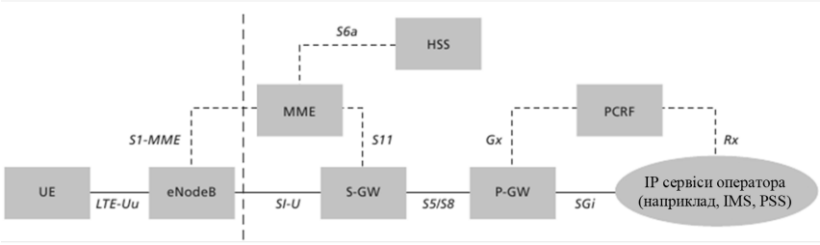


Рис. 2.13 Мережні елементи EPS [31]

Evolved Packet System (EPS) надає користувачеві IP підключення до Packet Data Network (PDN) для доступу в Інтернет, а також для запуску сервісів, таких як передача голосу по IP (Voice over IP – VoIP). EPS канал (bearer), як правило, асоціюється з якістю обслуговування (Quality of Service – QoS). Множинні канали можуть бути встановлені для користувача, для того, щоб забезпечити різні потоки QoS, а також підключення до різних PDN. Наприклад, користувач може здійснювати голосовий виклик (VoIP), в той же час виконуючи перегляд веб-сторінок або здійснюючи FTP-завантаження. VoIP канал забезпечить необхідну QoS для голосового виклику, в той час як канал з «найкращими зусиллями» буде цілком відповідним для перегляду веб-сторінок або FTP-сесії. Мережа повинна також забезпечити достатню безпеку і конфіденційність для користувача і захист мережі від несанкціонованого використання.

Це досягається за допомогою декількох мережевих елементів EPS, які мають різні ролі. На рис. 2.13 показана загальна архітектура мережі, включаючи мережеві елементи і стандартизовані інтерфейси. На високому рівні мережа складається з базової мережі (Core Network) (EPC) і мережі доступу E-UTRAN. У той час як базова мережа складається з безлічі логічних вузлів, мережа доступу складається з по суті тільки одного вузла, базової станції Evolved NodeB (eNodeB), який підключається до абонентського

обладнання (UE). Кожен з цих елементів мережі з'єднаний за допомогою інтерфейсів, які стандартизовані для того, щоб дозволити функціональну сумісність різних постачальників. Це дає операторам мережі можливість використовувати різні мережеві елементи від різних постачальників. Насправді, мережеві оператори можуть вибрати в їх фізичних реалізаціях розділити або об'єднати ці логічні мережеві елементи в залежності від комерційних міркувань. Функціональний розподіл між EPC і E-UTRAN показан на рис. 2.14.

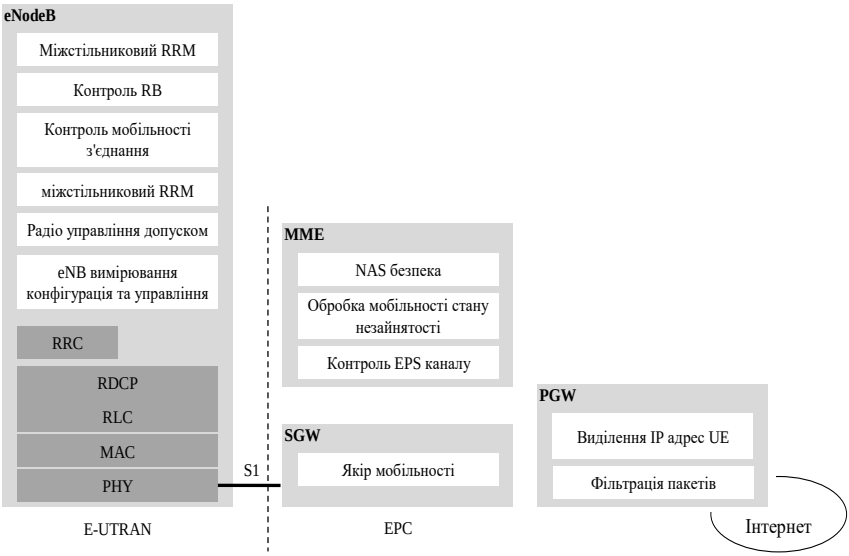


Рис. 2.14 Функціональний розподіл між E-UTRAN та EPC [31]

Базова мережа (яка називається EPC в LTE) відповідальна за загальний контроль над UE і встановлення каналів. Так як EPS забезпечує тільки шлях каналу певної QoS, управління мультимедійними додатками, такими як VoIP, забезпечується IP мультимедійною підсистемою (IP Multimedia Subsystem – IMS), яка, вважається, розташовується поза самої EPS.

Логічні вузли базової мережі показані на рис. 2.14 і більш докладно описані нижче [31]:

- PCRF – PCRF відповідає за прийняття рішень контролю політики, а також за управління тарифікацією на основі потоку в функції впровадження політики управління (Policy Control Enforcement Function – PCEF), що знаходиться в PGW. PCRF забезпечує авторизацію QoS (ідентифікатор класу QoS (QCI) і швидкість передачі бітів), що вирішує, як певний потік даних буде оброблятися в PCEF і гарантує, що це відповідає профілю підписки користувача.

- HSS – домашній абонентський сервер містить дані підписки користувачів SAE, такі як EPS-підписний профіль QoS і будь-які обмеження доступу для роумінгу. Він також містить інформацію про PDN, до яких користувач може підключитися. Це може бути у вигляді імені точки доступу (Access Point Name – APN) (яке являє собою позначку згідно з угодами про присвоєння імен DNS, що описує точку доступу для PDN) або адреси PDN (зазначаючи IP адресу/адреси підписки). Крім того, HSS має динамічну інформацію, таку як ідентифікатор MME, до яких користувач в даний момент підключений або зареєстрований. HSS також може інтегрувати центр аутентифікації (Authentication Center – AUC), який генерує вектори для ключів аутентифікації і безпеки.
- PGW – PDN Шлюз відповідає за розподіл IP адрес для UE, а також за забезпечення дотримання QoS і тарифікацію на основі потоку відповідно до правил з PCRF. Він відповідає за фільтрацію IP пакетів користувача низхідної лінії зв'язку в різні канали на основі QoS. Це виконується на основі шаблонів потоку трафіку (Traffic Flow Templates – TFTs). P-GW виконує впровадження QoS для каналів з гарантованою швидкістю передачі даних (Guaranteed Bit Rate – GBR). Він також служить як якір мобільності для взаємодії з не-3GPP технологіями, такими як мережі CDMA2000 і WiMAX.
- SGW – всі IP пакети користувача передаються через Обслуговуючий Шлюз, який служить локальним якорем мобільності для каналів передачі даних, коли UE переміщається між eNodeB. Він також зберігає інформацію про канали, коли UE знаходиться в стані очікування (відомий як «EPS управління з'єднанням – IDLE» [ECM-IDLE]) і тимчасово буферизує даних низхідної лінії в той час як MME ініціює пошуковиц виклик UE для повторного встановлення каналу. Крім того, S-GW виконує деякі адміністративні функції в гостьовій мережі, такі як збір інформації для тарифікації (наприклад, обсяг даних, переданих або отриманих від користувача) і законне перехоплення. Він також служить в якості якоря мобільності для взаємодії з іншими технологіями 3GPP, такими як GPRS і UMTS.
- MME – Mobility Management Entity (MME) є вузлом управління, який обробляє сигналізацію між UE і базовою мережею. Протоколи, що знаходяться між UE і базовою мережею відомі як протоколи Non Access Stratum (NAS).

Основні функції, підтримувані MME можуть бути класифіковані як:

- функції, пов'язані з управлінням каналу – включає в себе встановлення, підтримку і звільнення каналів і обробляється рівнем управління сесією в протоколі NAS;
- функції, пов'язані з управлінням з'єднанням –включає в себе встановлення з'єднання і безпеку між мережею і UE і обробляється рівнем з'єднання або управління мобільністю в рівні протоколу NAS.

Віртуалізовані мережеві функції (VNF), наприклад, S/P-GW, MME, можуть масштабуватися незалежно один від одного відповідно до їх конкретних потреб в ресурсах, наприклад може виникнути ситуація, коли необхідно збільшити ресурси площини користувача, не впливаючи на площину управління, і навпаки. Крім того, VNF, які мають справу з площиною даних може знадобитися різна кількість ресурсів NFV інфраструктури, ніж тим VNF, що займаються тільки сигналізацією.

Різні сценарії можуть існувати, де, наприклад, весь EPC віртуалізовано в одній точці присутності інфраструктури NFV (NFVI) або тільки деякі мережеві функції віртуалізовано.

Для досягнення безперервності обслуговування і доступності послуг бажаних оператором, надійність повинна бути забезпечена і в площині управління, і в площині даних. Оскільки віртуалізація забезпечує відв'язку мережевої функції від нижчерозташованого обладнання, проектування нових схем відмовостійкості стає можливим за рахунок використання портативності екземплярів VNF. На рис. 2.15 [28] показано віртуалізовані EPC і IMS, де віртуалізовані S/P-GW і функції IMS оброблюють PDN з'єднання і IMS сесії відповідно. Коли динамічне переміщення цих екземплярів VNF виконується через перевантаження віртуальної машини або відмову в автоматичному режимі або на вимогу, переміщення керованих сесій і/або з'єднань повинні бути оброблені відповідним чином для досягнення бажаних оператором безперервності обслуговування і доступності послуг.

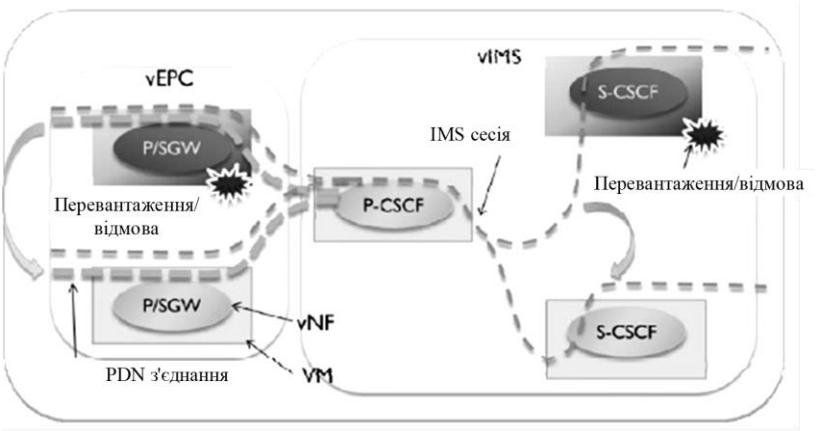


Рис. 2.15 Переміщення VNF

Нижче перераховані проблеми високого рівня, які повинні бути прийняті до уваги при визначенні конкретних рішень для цього випадку використання [28]:

1. Масштабування ресурсів: Розширення і згортання мережевих ресурсів віртуалізованих мережевих функцій.
2. Обізнаність в сервісах: Сервіс-обізнаний розподіл ресурсів для мережевих функцій.
3. Прозорість віртуалізації для сервісів: Сервіси використовуючи мережеву функцію не повинні знати, чи це віртуальна функція або не-віртуалізована.
4. Прозорість віртуалізації для мережевого контролю і управління: Площина мережевого контролю та управління не повинна бути обізнаною в тому, чи функція віртуалізована чи ні.

5. Управління станом: управління станом мережі і мережевої функції під час перенесення мережевої функції, реплікації і масштабування ресурсів.

6. Моніторинг/виявлення помилок/діагностика/відновлення: Відповідний механізм для моніторингу/виявлення помилок/діагностики/відновлення всіх компонентів і їх станів після віртуалізації, наприклад, екземплярів VNF, апаратного забезпечення, гіпервізора.

7. Доступність сервісу: Досягнення такого ж рівня доступності сервісів для віртуалізованої мобільної опорної мережі з кінця в кінець, як в не віртуалізованих мережах з меншими витратами.

8. Механізм поділу управління трафіком: Ідентифікація/поділ трафіку даних і управління для не-віртуалізованих і віртуалізованих мобільних базових мереж.

9. Вплив на функції: Зведення до мінімуму впливу на існуючі не-віртуалізовані мережеві функції та підтримка мережевих операційних систем.

EPCaaS можна розглядати як хмарну реалізацію 3GPP EPC архітектури, що може надаватися на вимогу [25], за допомогою архітектурної платформи, описаної в цій роботі.

2.3 Якість обслуговування та EPS канали

Розглянемо вимоги до якості обслуговування в мережах LTE [31].

У типовому випадку, кілька прикладних програм можуть бути запущені на UE в будь-який момент, кожна з яких має різні вимоги якості обслуговування. Наприклад, UE може здійснювати виклику VoIP в той же час переглядаючи веб-сторінку або завантажуючи FTP-файл. VoIP має більш жорсткі вимоги до QoS в термінах затримки і джиттера затримки, ніж перегляд веб-сторінок і FTP, в той час як останній вимагає набагато нижчу частоту втрати пакетів. Для того, щоб підтримувати кілька вимог QoS, встановлюються різні канали в рамках EPS, кожен з яких пов'язаний з QoS.

У широкому сенсі, канали можуть бути класифіковані на дві категорії, основуючись на характері QoS, які вони забезпечують:

1. Канали з мінімальною гарантованою швидкістю передачі даних (Guaranteed Bit Rate – GBR), які можуть бути використані для прикладних програм, таких як VoIP. Вони мають відповідне значення GBR, для якого виділені ресурси передачі постійно виділяються (наприклад, за допомогою функції управління доступом в eNodeB) при встановленні або модифікації каналу. Швидкості передачі вищі, ніж GBR можуть бути доступніо для GBR каналу при наявності ресурсів. У таких випадках, максимальна швидкість передавання даних (Maximum Bit Rate – MBR), параметр, який також може бути пов'язаний з GBR каналом, встановлює верхню межу швидкості передачі в бітах, що можна очікувати від GBR каналу.

2. Не-GBR канали, які не гарантують будь-якої конкретної швидкості передачі даних. Вони можуть бути використані для таких прикладних програм, як перегляд веб-сторінок або FTP-передача. Для цих каналів ніякі ресурси смуги пропускання не виділяються постійно каналу.

У мережі доступу, eNodeB відповідає за те, щоб забезпечити необхідну QoS для каналу передачі через інтерфейс радіозв'язку. Кожен канал має пов'язаний з ним ідентифікатор класу якості обслуговування (QoS Class Identifier – QCI), а також пріоритет розподілу і утримання (Allocation and Retention Priority – ARP).

Кожен QCI характеризується пріоритетом, граничним допустимим значенням затримки пакетів і прийнятним рівнем втрат пакетів. Мітка QCI для каналу визначає, як він обробляється в eNodeB. Тільки десяток таких QCI було стандартизовано, так що всі вендори можуть мати однакове розуміння відповідних характеристик сервісів і тим самим забезпечити відповідну обробку, включаючи управління чергою, стратегію політик та перетворення.

Це гарантує, що оператор LTE може очікувати однакову поведінку обробки трафіку по всій мережі, незалежно від виробників eNodeB обладнання. Набір стандартизованих QCI і їх характеристик (з яких PCRF в EPS може вибирати) представлений в таблиці 2.1 (з розділу 6.1.7 в [32]). Таблиця QCI визначає значення для обробки пріоритетів, прийнятної затримки і кількості втрачених пакетів для кожної мітки QCI.

Таблиця 2.1

Стандартизовані QCI для LTE

QCI	Тип ресурсу	Пріоритет	Допустима затримка пакету (мс)	Рівень пакетних помилок	Приклади сервісів
1	GBR	2	100	10^{-2}	Розмовний голос
2	GBR	4	150	10^{-3}	Розмовне відео (живий стримінг)
3	GBR	5	300	10^{-6}	Не-розмовне відео (буферизований стримінг)
4	GBR	3	50	10^{-3}	Ігри реального часу
5	Не-GBR	1	100	10^{-6}	IMS сигналізація
6	Не-GBR	7	100	10^{-3}	Голос, відео (живий стримінг), інтерактивні ігри
7	Не-GBR	6	300	10^{-6}	Відео (буферизований стримінг)
8	Не-GBR	8	300	10^{-6}	TCP (наприклад, WWW, e-mail), чат, FTP, p2p поділ файлів, тривале відео та інші
9	Не-GBR	9	300	10^{-6}	

Пріоритет і гранична затримка пакетів (і до певної міри прийнятний коефіцієнт втрати пакетів) з мітки QCI визначають конфігурацію режиму Radio Link Control (RLC) (дивіться розділ 4.3.1 з [33]) і як планувальник в MAC обробляє пакети, що передаються через канал (наприклад, з точки зору політики планування, політики управління чергами та політики формування швидкості). Наприклад, пакет з більш високим пріоритетом, буде плануватись перед пакетом з більш низьким пріоритетом. Для каналів передачі з низьким прийнятним рівнем втрат, режим з підтвердженням (Acknowledged Mode) може бути використаний в рівні протоколу RLC, щоб гарантувати, що пакети успішно доставлені через радіоінтерфейс.

ARP каналу використовується для контролю допустимості викликів – тобто, щоб вирішити, чи слід встановлюватися запитуваний канал в разі радіо перевантаження. Він також керує пріоритезацією каналу для захоплення по відношенню до нового запиту встановлення каналу. Після успішного встановлення, ARP каналу не робить ніякого впливу на обробку пересилання пакетів на рівні каналу (наприклад, для планування і управління швидкістю). Така обробка пересилання пакетів повинна виключно визначатися іншими параметрами QoS рівня каналу, такими як QCI, GBR і MBR.

2.4 Віртуалізація Мережевих Функцій (NFV)

2.4.1 Високорівнева платформа NFV

NFV передбачає реалізацію мережевих функцій як сутностей програмного забезпечення, які працюють на Інфраструктурі NFV (NFVI). На рис. 2.16 [34] показано платформу NFV високого рівня. Таким чином, три основні робочі області визначені в NFV [35]:

- Віртуалізована Мережева Функція (VNF), як програмна реалізація функції мережі, яка здатна працювати на NFVI;
- Інфраструктура NFV (NFVI), включаючи різноманітність фізичних ресурсів і як вони можуть бути віртуалізовані. NFVI підтримує виконання VNF;
- Управління і Оркестровка NFV (NFV MANO), що охоплює оркестровку і управління життєвим циклом фізичних і/або програмних ресурсів, що підтримують віртуалізацію інфраструктури, та управління життєвим циклом VNF.

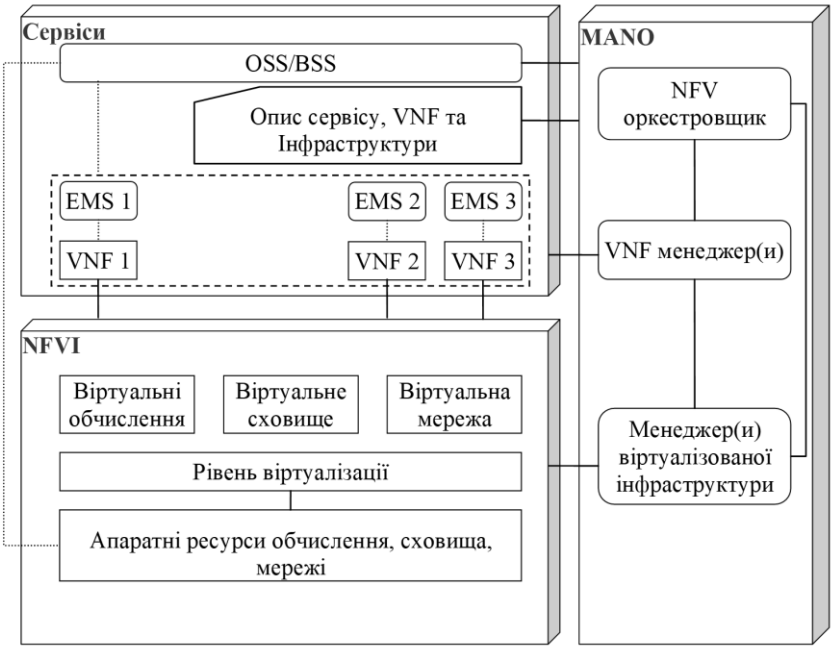


Рис. 2.16 Високорівнева платформа NFV [34]

Згідно з платформою MANO від ETSI, NFV MANO забезпечує функціональні можливості, необхідні для управління VNF, і пов'язаних з цим операцій, такі як конфігурація VNF та інфраструктури, на яких ці функції працюють. Вона включає в себе оркестровку і управління життєвим циклом фізичних і/або програмних ресурсів, що підтримують віртуалізацію інфраструктури, а також управління життєвим циклом VNF. Вона також включає в себе бази даних, які використовуються для зберігання моделей інформації і даних, які визначають як властивості розгортання, так і властивості життєвого циклу функцій, сервісів і ресурсів. NFV MANO фокусується на всіх завданнях управління віртуалізації необхідних платформі NFV. Крім того, платформа визначає інтерфейси, які можуть використовуватися для зв'язку між різними компонентами NFV MANO, а також координації з традиційними системами керування мережею, такими як системи підтримки операцій (Operations Support System – OSS) і системи підтримки бізнесу (Business Support System – BSS), для того, щоб забезпечити керування як VNF, так і функціями, які працюють на традиційному обладнанні [18].

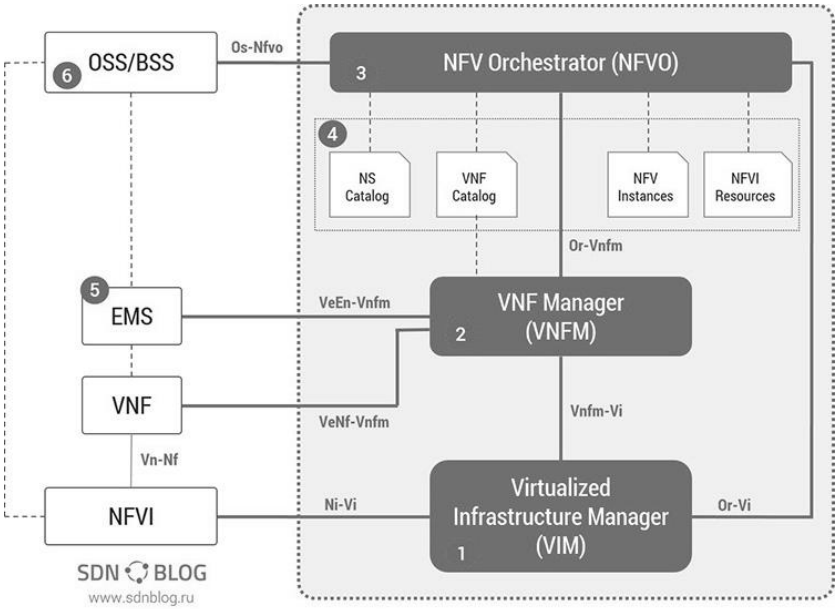


Рис. 2.17 ETSI-NFV архітектура [36]

На рис. 2.17 [36] MANO – це виділений блок праворуч, що складається з трьох менеджерів і групи репозиторіїв (4):

1. Virtualized Infrastructure Manager (VIM).
2. VNF Manager (VNFM).
3. NFV Orchestrator (VNFO).
4. Репозиторії NS Catalog, VNF Catalog, NFV Instances і NFVI Resources.

На додаток до чотирьох блоків всередині MANO, на схемі присутні два блоки – традиційні Element Management System (EMS) і OSS/BSS. Хоча останні два блоки не є частиною MANO, вони обмінюються з нею інформацією.

Virtualized Network Function (VNF) – це сама віртуальна функція (наприклад, віртуальний DPI, Firewall, Router і ін.).

NFV Infrastructure (NFVI) визначає інфраструктуру – фізичні (сервери, диск та ін.), віртуальні ресурси (віртуальні машини) і гипервизор.

Які функції виконує VIM:

- управління життєвим циклом ресурсів в домені NFVI, включаючи створення, управління станом і видалення віртуальних машин;
- зберігання бази віртуальних машин і асоційованих з ними фізичних ресурсів;
- моніторинг продуктивності, виявлення відмов обладнання, програмного забезпечення та віртуальних ресурсів;

- надає інтерфейс (Northbound API) іншим системам управління для надання інформації про фізичні та віртуальні ресурси.

VNFM управляє віртуальної мережевої функцією (VNF). Він як VIM для NFVI, тільки для VNF.

Які функції він виконує:

- управління життєвим циклом VNF - створення, управління станом і видалення VNF Instances (самі VNF встановлені на віртуальних машинах, які знаходяться під управлінням VIM);
- конфігурація VNF, моніторинг і статистика, метрики продуктивності і управління політиками безпеки;
- управління розміром VNF – збільшення/зменшення ресурсів CPU, пам'ять, диск.

NFV Orchestrator (NFVO) виконує наступні функції:

- Оркестрація ресурсів. NFVO управляє виділенням і резервуванням ресурсів і дає дозволи на їх використання або в рамках одного або декількох доменів NFVI. Управління здійснюється через відповідні VIM за допомогою API, а не за допомогою прямої взаємодії з ресурсами NFVI.
- Оркестрація сервісів. NFVO відповідає за створення призначених для користувача сервісів з декількох VNF (які можуть управлятися різними VNF менеджерами). Оркестрація сервісів відбувається наступним чином:
    - NFVO створює сервіс за допомогою взаємодії з відповідним VNFM, без необхідності прямої взаємодії з VNF;
    - NFVO може звертатися до VNFM для отримання будь-якої інформації про VNF;
    - зберігає топологію створених ланцюжків сервісів (англ. VNF Forwarding Graphs або Service Chains).

NFVO «склеює» воєдино інші функції MANO і надає уніфікований інтерфейс для управління ресурсами і сервісами в рамках мультивендорної платформи.

Репозиторії в NFV MANO зберігають різну інформацію, в залежності від їх типу.

VNF Catalog – репозиторій всіх використовуваних VNF дескрипторів (VNFD). VNF Descriptor - шаблон, який визначає параметри VNF і вимоги до впровадження та експлуатації. Використовується VNF менеджером в процесі створення VNF і управління життєвим циклом віртуальної мережевої функції. Інформація, яку надає VNF дескриптор, також використовується NFVO для оркестрації мережевих сервісів і інфраструктурних ресурсів.

Network Services (NS) Catalog – список використовуваних для користувача мережевих сервісів. Шаблон мережевого сервісу надає опис взаємодії VNF між собою (Наприклад, характеристики віртуального каналу).

NFV Instances – список даних про прив'язку користувальницьких мережевих сервісів і відповідних VNF. Наприклад, два різних клієнта придбали віртуальний маршрутизатор. Щоб відрізняти віртуальний маршрутизатор клієнта 1 від віртуального маршрутизатора клієнта 2, необхідний список з прив'язкою ID маршрутизаторів до конкретного клієнта.

NFVI Resources – репозиторій для зберігання використовуваних ресурсів. Необхідний для створення нових VNF сервісів або масштабування існуючих.

EMS відповідає за конфігурацію VNF, моніторинг і збір статистики, метрики продуктивності і управління політиками безпеки. Хоча VNFM робить те ж саме, однак, EMS виконує ці функції за допомогою пропрієтарних/закритих інтерфейсів взаємодії з VNF на відміну від VNFM. Потрібен відкритий «інтерфейс» взаємодії (Ve-Vnfm-em), в MANO званий Reference Point.

Взаємодія зовнішньої EMS з VNFM може знадобитися для отримання інформації щодо віртуальних ресурсів, асоційованих з даною мережевою функцією (VNF).

OSS/BSS – IT-системи оператора. Природно, NFV повинна працювати в тісній взаємодії з цими системами.

В принципі можна було розширити функціональність існуючих OSS/BSS систем і «навчити» їх управляти VNF і NFVI безпосередньо. Але в більшості випадків все впиралося б в реалізацію від конкретного виробника. NFV позиціонується як відкрита платформа, і управління її елементів повинно здійснюватися через відкриті інтерфейси (як в MANO).

Проте, існуючі OSS/BSS можуть додати бізнес-переваг до NFV MANO шляхом пропозиції додаткових функцій, які не підтримуються конкретною реалізацією архітектури. Для цього в архітектурі передбачений ще один Reference Point (Or-Ma-NFVO), що описує взаємодію MANO з OSS/BSS системами.

2.4.2 Функції та сервіси віртуальної мережі

Мережева функція є функціональним блоком в межах мережевої інфраструктури, яка має чітко визначені зовнішні інтерфейси і чітко визначену функціональну поведінку. Прикладами мережевих функцій є елементи в домашній мережі, наприклад абонентський шлюз (Residential Gateway – RGW); і традиційні мережеві функції, наприклад, DHCP-сервери, брандмауери і т.д. Таким чином, VNF є реалізацією мережевої функції, яка розгорнута на віртуальних ресурсах, таких як віртуальна машина. Один VNF може складатися з декількох внутрішніх компонентів і, отже, він може бути розгорнутий на декількох віртуальних машинах, і в цьому випадку кожна віртуальна машина містить один з компонентів VNF. Сервіс є пропозицією від TSP, що складається з однієї або декількох мережевих функцій. У випадку NFV, мережеві функції, які утворюють сервіс, віртуалізуються і розгортаються на віртуальних ресурсах, таких як віртуальна машина. Проте, з точки зору користувачів, сервіси, що працюють на функціях спеціалізованого обладнання або на віртуальних машинах, повинні мати однакову продуктивність. Кількість, тип і порядок VNF, що створюють його, визначаються функціональною і поведінковою специфікацією сервісу. Таким чином, поведінка сервісу залежить від складових VNF [18].

Мережевий сервіс архітектурно може розглядатися як граф мережевих функцій, з'єднаних між собою за допомогою відповідної мережевої інфраструктури. Ці мережеві функції можуть бути реалізовані в мережі одного оператора або у взаємодії між різними операторами мереж. Поведінка нижчерозташованої мережевої функції вносить вклад у поведінку сервісу вищого рівня. Таким чином, поведінка мережевого сервісу являє собою поєднання поведінки її складових функціональних блоків, які можуть включати в себе

окремі мережеві функції, набори мережевих функцій, графи мережевих функцій і/або мережу інфраструктури.

Кінцеві точки і мережеві функції мережевого сервісу представлені у вигляді вузлів і відповідають пристроям, прикладним програмам і/або прикладним програмам фізичних серверів. Граф мережевих функцій може містити вузли мережевих функцій, з'єднаних логічними каналами, які можуть бути однонаправленими, двонаправленими, багатонаправленими та/або широкомовними. Простим прикладом графа є ланцюг мережевих функцій. Приклад такого мережевого сервісу з кінця в кінець може включати в себе смартфон, бездротову мережу, брандмауер, балансувальник навантаження і набір серверів мережі доставки контента. Область діяльності NFV знаходиться всередині ресурсів, якими володіє оператор. Таким чином, пристрій користувача, наприклад, мобільний телефон знаходиться поза зоною дії, так як оператор не може управляти ним. Проте, віртуалізація і мережевий хостинг функцій абонента можливі і знаходяться в рамках NFV (наприклад, див. випадки використання Віртуальної Платформи як Сервіс (Virtual Network Platform as a Service – VNPaaS) і віртуалізацію домашнього середовища в GS NFV 001 [28]).

Рис. 2.18 ілюструє представлення мережевого сервісу з кінця в кінець, що включає в себе другий вкладений граф мережевих функцій, як показано блоками вузлів мережевих функцій в середині рисунку, з'єднаних логічними каналами. Кінцеві точки підключаються до мережевих функцій через мережеву інфраструктуру (дротову або бездротову), в результаті чого ми бачимо логічний інтерфейс між кінцевою точкою і функцією мережі. Ці логічні інтерфейси представлені на рисунку пунктиром. На рис. 2.18, зовнішній мережевий сервіс з кінця в кінець складається з кінцевої точки A, внутрішнього графа мережевих функцій і кінцевої точки B, в той час як внутрішній граф мережевих функцій складається з мережевих функцій NF1, NF2 і NF3. Вони з'єднані між собою за допомогою логічних каналів, що надаються мережею інфраструктури 2 [35].

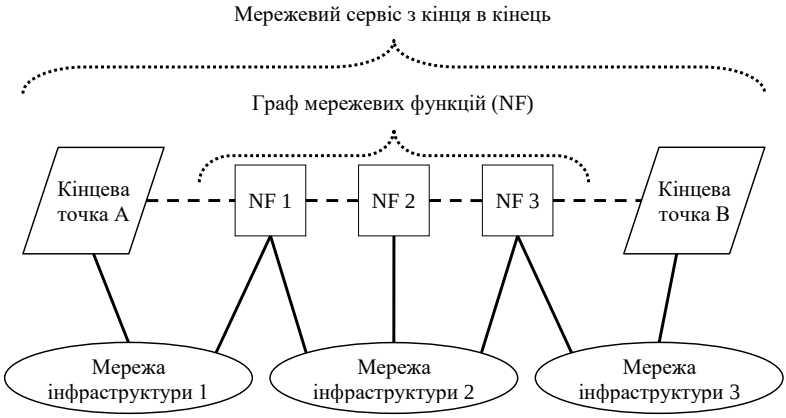


Рис. 2.18 Представлення мережевого сервісу з кінця в кінець у вигляді графа [35]

2.5 Хмарні обчислення

2.5.1 Загальна характеристика

Організація процесу надання послуг великій кількості користувачів, яка оцінюється в мільйони, має спільні риси. Так, все більше компаній користуються послугами хмарних сервісів, які забезпечені технічними ресурсами та можуть надавати інфраструктури як сервіс [37].

Хмарні обчислення (Cloud Computing) – це модель забезпечення повсюдного та зручного доступу на вимогу через мережу до спільного пулу обчислювальних ресурсів, що підлягають налаштуванню (наприклад, до комунікаційних мереж, серверів, засобів збереження даних, прикладних програм та сервісів), і які можуть бути оперативно надані та звільнені з мінімальними управлінськими затратами та зверненнями до провайдера [38].

Для того, щоб розмістити віртуальну машину, фізична машина повинна надати всі ресурси, які вимагає віртуальна машина, в тому числі використання процесора, пам'ять, сховище та пропускну здатність мережевої карти [33].

Модель хмарних обчислень складається з п'яти ключових характеристик і трьох моделей обслуговування [40]. Коротко представимо їх далі.

1) Основні характеристики хмарних обчислень:

1. Самообслуговування на-вимогу. Споживач може самостійно керувати обчислювальними можливостями, такими як час сервера і мережеве сховище, в міру необхідності автоматично без потреби у взаємодії людини з кожним сервіс провайдером.

2. Широкий доступ до мережі. Можливості (наприклад, обчислювальні ресурси, ємності зберігання) доступні через мережу і отримати до них доступ можна через стандартні механізми, які використовують різнорідні платформи тонкого або товстого клієнта (наприклад, мобільні телефони, планшети, ноутбуки і робочі станції).

3. Об'єднання ресурсів у пул. Обчислювальні ресурси провайдера об'єднують у пул для обслуговування багатьох споживачів з використанням мультитенантної моделі, з різними фізичними та віртуальними ресурсами, які динамічно призначаються і перепризначаються відповідно до споживчого попиту.

4. Швидка еластичність. Потужності можуть еластично надаватися і звільнятися, в деяких випадках автоматично, щоб швидко масштабуватися назовні і всередину пропорційно до вимог.

5. Виміряне обслуговування. Хмарні системи автоматично керують та оптимізують використання ресурсів за рахунок використання можливості вимірювання на певному рівні абстракції, відповідному типу обслуговування (наприклад, зберігання, обробки, пропускної здатності, і активних облікових записів користувачів).

2) Моделі обслуговуваня хмарних обчислень: три сервісні моделі хмарних обчислень показані на рис. 2.19, і визначені нижче [41]:

1. Програмне забезпечення як сервіс (Software as a Service – SaaS). Користувач має можливість використовувати прикладні програми провайдерів, що працюють на хмарній інфраструктурі. Прикладні програми є доступними з різних клієнтських пристроїв або через тонкий клієнтський інтерфейс, такий як веб-браузер (наприклад, веб-пошта) або інтерфейс програмування.

2. Платформа як сервіс (Platform as a Service – PaaS). Користувач має можливість розгорнути у інфраструктурі хмари створені споживачем або придбані прикладні програми, створені з використанням мов програмування, бібліотек, сервісів і інструментів, підтримуваних провайдером.

3. Інфраструктура як сервіс (Infrastructure as a Service – IaaS). Користувач має можливість управляти ресурсами обробки, зберігання, мереж та іншими фундаментальними обчислювальними ресурсами, де споживач має можливість розгорнути і запустити довільне програмне забезпечення, яке може включати в себе операційні системи і прикладні програми.

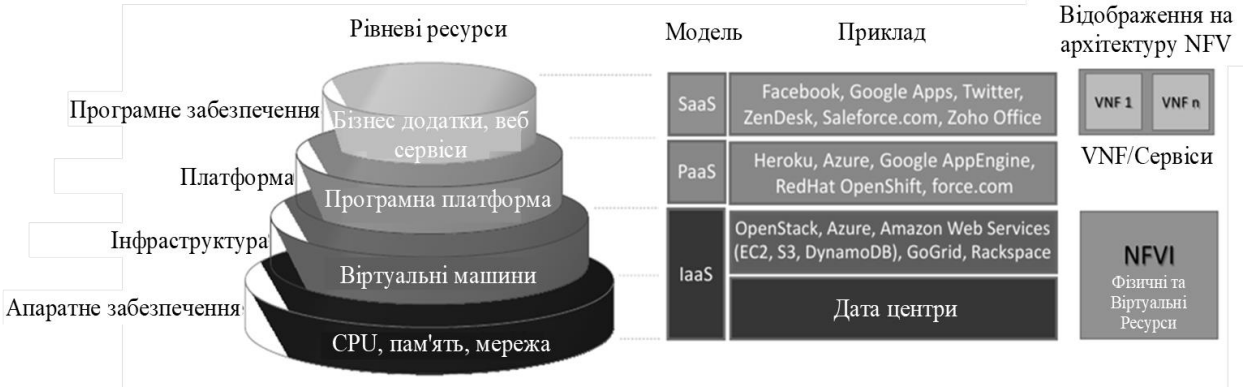


Рис. 2.19 Сервісні моделі хмарних обчислень та їх відображення на частину еталонної архітектури NFV

2.5.2 Взаємозв'язок між хмарними обчисленнями та NFV

Розгортання мережевих функцій в хмарі, ймовірно змінить всі аспекти, пов'язані з тим як сервіси і прикладні програми розробляються і поставляються. У той час як продовжуються робити відносно мережевих хмар і взаємодії між хмарами [41], [42], телекомунікаційні мережі відрізняються від середовища хмарних обчислень принаймні трьома способами: (1) навантаженя площини даних в телекомунікаційних мережах мають високий тиск на продуктивність, (2) топології телекомунікаційних мереж накладають жорсткі вимоги на мережу і потребу в глобальному погляді на мережу для керування [43], (3) телекомунікаційна галузь вимагає масштабованості, доступності п'ять дев'яток та надійності. У традиційних телекомунікаційних мережах, ці характеристики забезпечуються інфраструктурою сайту. Якщо NFV засновується на хмарних обчисленнях, ці функції повинні бути відтворені за допомогою хмарної інфраструктури таким чином, що вони можуть бути оркестровані, а оркестрові функції можуть бути доступні через відповідні абстракції, а також бути поєднані з розширеною підтримкою для виявлення і простежуваності [44]. У зв'язку з цим слід підкреслити, що NFV поставе більше задач, ніж просто перенесення функцій мережі операторського класу в хмару. Існує необхідність адаптувати хмарні середовища, щоб отримати поведінку операторського класу [43]. У Таблиці 2.2 підсумовується зв'язок між NFV для телекомунікаційних мереж і хмарними обчисленнями [18].

Таблиця 2.2

Порівняння NFV у телекомунікаційних мережах та хмарних обчислень

Проблема	**NFV (телекомунікаційні мережі)**	**Хмарні обчисленя**
підхід	абстракція сервісу/функцій	абстракція обчислень
формалізація	ETSI NFV група стандартів промисловості	робоча група управління хмарами DMTF [45]
затримка	очікування низької затримки	деяка затримка допустима
інфраструктура	гетерогенний транспорт (оптичний, безпровідний, Ethernet)	гомогенний транспорт (Ethernet)
протокол	багато протоколів управління (наприклад OpenFlow [46], SNMP [47])	OpenFlow
надійність	суворі вимоги доступності 5 Дев'яток [48]	менш суворі вимоги надійності[49]
регуляція	суворі вимоги, наприклад NEBS [50]	різноманітні та змінювані

2.5.3 NFV, SDN та хмарні обчислення

Щоб підвести підсумок взаємозв'язку між NFV, SDN і хмарними обчисленнями, використовуємо рис. 2.20 (варто зауважити, що OpenFlow не єдиний протокол SDN. Точно так, OpenStack не єдина платформа хмарних обчислень. Причиною наведення

лише цих двох на рис. 2.20 є те, що вони отримали більше уваги в цілому, а також і по відношенню до NFV) [18]. Зауважимо, що кожне з цих полів є абстракцією різних ресурсів: обчислювальних для хмарних обчислень, мережевих для SDN, і функцій для NFV. Переваги, які отримуються від кожного з них схожі: гнучкість, зниження витрат, динамічність, автоматизація, масштабування ресурсів і т.д.

Питання не в тому, чи будуть мережеві функцій перенесені в хмару, так як це насправді загальна ідея NFV. Воно в тому чи це буде хмара публічна, як Amazon, або якщо телекомунікаційні сервіс провайдери віддадуть перевагу використанню приватних хмар, розподілених по всій їх інфраструктурі. У будь-якому випадку, необхідно створити хмару операторського класу з точки зору продуктивності, надійності, безпеки, зв'язку між функціями і т.д.

З іншого боку, цілі NFV можуть бути досягнуті за допомогою механізмів, що не відносяться до SDN, і спираються на методи, що використовуються в даний час в багатьох дата центрах. Проте, підходи, що спираються на поділ площин управління і передачі даних як пропонується SDN, можуть підвищити продуктивність, спростити сумісність з існуючими реалізаціями, а також спростити процедури експлуатації та технічного обслуговування. Таким же чином, NFV здатний підтримувати SDN, надаючи інфраструктуру, на якій програмне забезпечення SDN може бути запущене. І, нарешті, сучасний варіант дата центру (хмари і їх аспект самообслуговування) спирається на автоматизовані системи керування, які можуть бути отримані з SDN та NFV. Зокрема, такі аспекти, як мережа як сервіс, балансування навантаження, брандмауер, VPN і т.д. всі працюють в програмному забезпеченні що запускається за допомогою API.

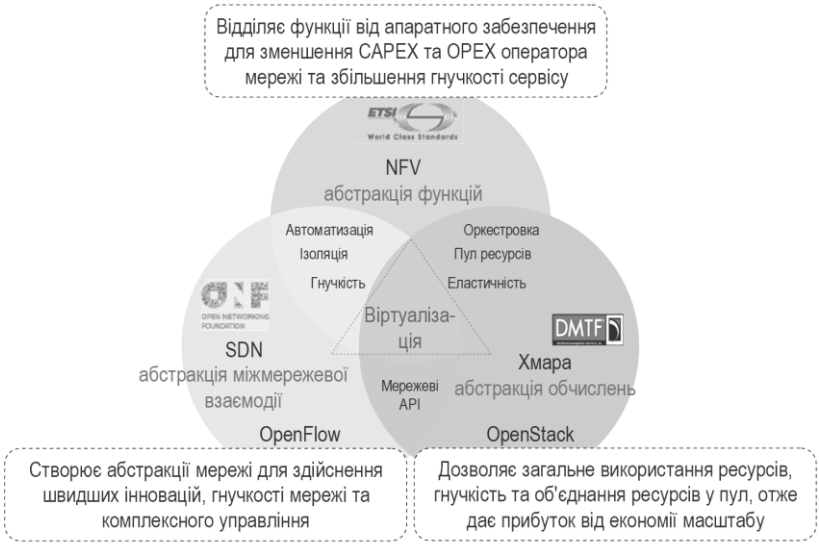


Рис. 2.20 Взаємозв'язок між NFV, SDN та хмарними обчисленнями

2.6 Опис мережевого сервісу

NFV дозволяє операторам масштабувати послуги мережі з більшою деталізацією та оперативністю, ніж сьогодні, коли продуктивність мережевого сервісу визначена статично на найбільший прогнозований пік, що веде до надлишковості ресурсів. Для цього автоматизація є ключовою. У пошуках цієї автоматизації інститут ETSI визначив еталонну архітектуру NFV, яка використовує шаблони, керовані моделями, що називаються NSD (NS descriptor (NSD)) для забезпечення роботи мережевих сервісів. Для операції масштабування, NSD визначається дискретний набір рівнів функціонування, серед яких екземпляр мережевого сервісу може бути змінений у розмірі упродовж життєвого циклу [51].

VNF Descriptor (VNFD) – це шаблон специфікації, що надається VNF Провайдером для опису вимог віртуальних ресурсів VNF. Він використовується функціями NFV MANO для визначення способу виконання операцій життєвого циклу VNF (наприклад, створення екемпляра) [52].

На рис. 2.21 [53] показано спрощений приклад дескриптора VNF (VNFD) та його зв'язок з VNF. Приклад на рис. 2.21 показує екземпляр VNF, який складається з 4 сутностей VNFC 3 різних типів: 'A', 'B' та 'C'. Кожен тип VNFC має власні вимоги до операційної системи (ОС) та середовища виконання (наприклад, віртуальної машини). Ці віртуальні ресурси та їхні вимоги до зв'язку описуються в елементах моделі даних, які формують VNFD. Окрім вимог до ресурсів, VNFD також містить однозначні посилання на файли двійкового коду VNF, скрипти, дані конфігурації тощо, які необхідні для управління NFV та функцій оркестровки для належного налаштування VNF.

Вимоги до ресурсів NFVI (наприклад, вимоги до з'єднання, пропускної спроможність, затримки тощо) не показані на рис. 2.21, але передбачається, що вони присутні в VNFD, а також в інших дескрипторах, які використовуються функціями керування та оркестрування NFV. VNFD може також вказувати правила розташування (наприклад, деякі екземпляри VNFC повинні розміщатися на віртуальних машинах, які надаються ресурсами, розташованими в одній стійці).

Функції керування та оркестрування NFV враховують всі доступні атрибути VNFD, щоб перевірити можливість створення екземпляру даного VNF, наприклад. перевіряють, які типи ресурсів потрібні для кожного екземпляра VNFC. Є кілька варіантів того, як можна створювати екземпляри окремих VNFC, а саме:

- повністю або частково завантажені контейнери віртуалізації;
- порожні контейнери для віртуалізації, підготовлені для завантаження.

Наприклад, якщо використовується останній варіант, то файл завантаження контейнера віртуалізації буде NULL. Потім відповідальністю функцій керування та оркестування VNF (наприклад, NFVO та VNF Manager) буде наказати VIM створити порожній контейнер віртуалізації з відповідним інтерфейсом SWA-5 (Vn-Nf), який готовий до використання.

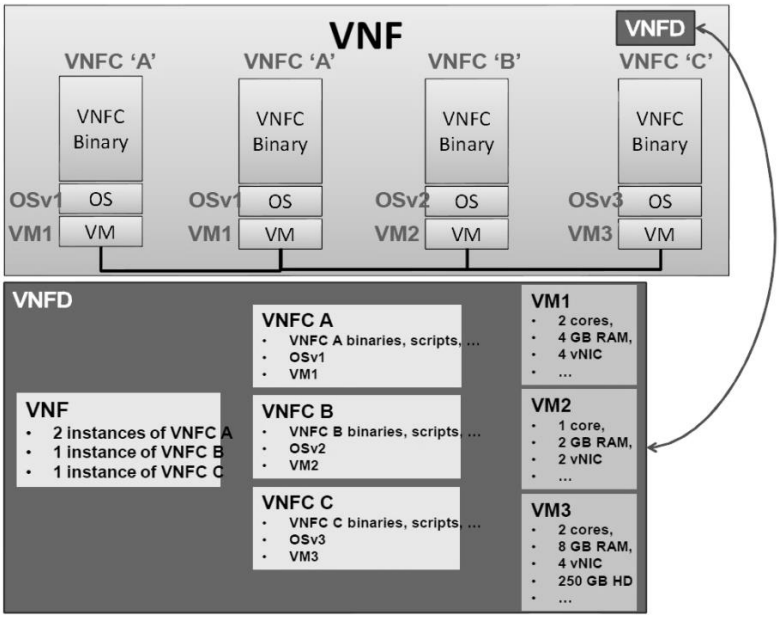


Рис. 2.21 Приклад елементів VNFD [53]

Таким чином, Дескриптор Мережевого Сервісу (Network Service Descriptor – NSD) – це шаблон розгортання, який містить машинно оброблювану інформацію, що використовується блоками MANO, для створення екземплярів мережевого сервісу (NS) та керування ними протягом усього життя. NSD побудований з набору атрибутів та інших дескрипторів (рис. 2.22) [51].

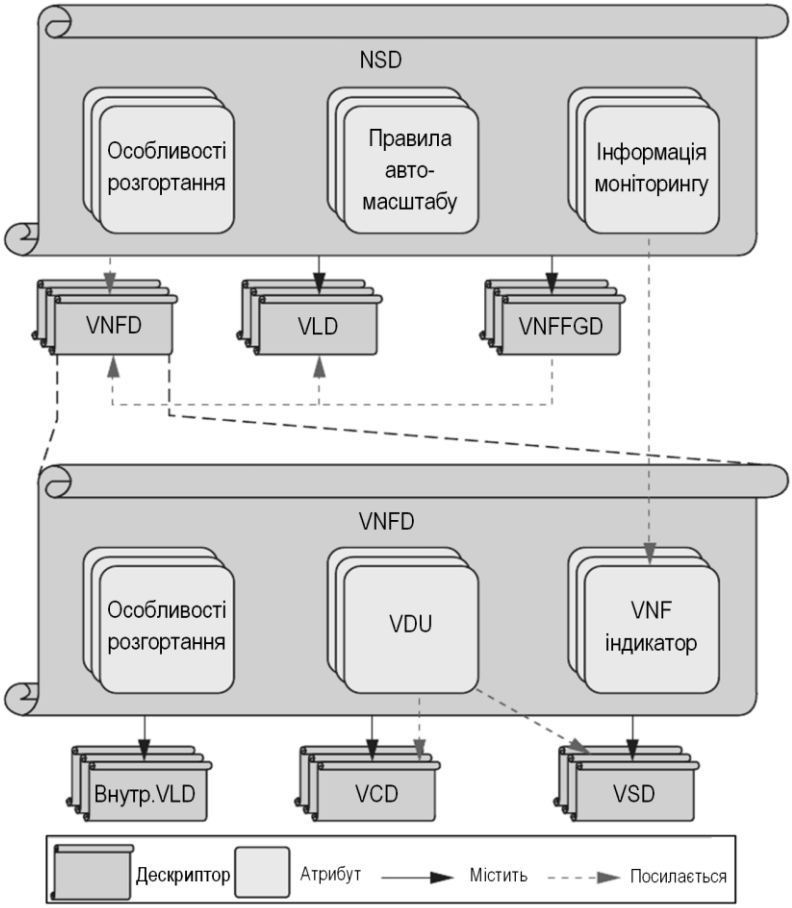


Рис. 2.22 Структура NSD. Показано дескриптори та атрибути найбільш релевантні для масштабування

Інформація моніторингу: визначає інформацію, яку слід відслідковувати для керування продуктивністю та несправностями NS. Ця інформація включає показники ефективності, пов'язані з ресурсами (на рівні NS/VNF) та індикатори VNF від складових VNF.

Правила автоматичного масштабування: містять правила, які дозволяють запускати дії по масштабуванню на екземплярі NS, коли умова, що містить інформацію моніторинга, не виконується. Інформаційна модель NFV дозволяє виражати ці правила як індивідуальні скрипти, що надаються під час створення екземпляру.

Особливості розгортання: описують конкретні конфігурації розгортання для NS.

Для опису розгортання та операційної поведінки компонентів NS, NSD містить і звертається до набору дескрипторів, включаючи дескриптори VNF, VL-дескриптори (VLDs) та дескриптори VNFFG (VNFFGDs) [54]. VNFD містить інформацію, необхідну для розгортання та функціонування екземплярів VNF. VLD надає інформацію про VL, включаючи конфігурації розгортання, доступні для створення екземпляру VL. Ці конфігурації визначаються за допомогою особливостей розгортання. Різні конфігурації забезпечують різні рівні продуктивності та надійності VNF. Нарешті, VNFFGD посилається на VNFDs та VLDs для опису топології.

З наведених дескрипторів розглянемо VNFD. NSD посилається на інформацію VNFD, яка є важливою для масштабування NS. Як і в NSD, VNFD також включає дескриптори та атрибути.

Дескриптори, що містяться в VNFD, містять детальний опис внутрішньої композиції VNF. Зокрема, VNFD включає в себе дескриптори віртуальних обчислень(virtual compute descriptors VCD), дескриптори віртуального зберігання (virtual storage descriptors VSD) та внутрішні VLD. Перші два вказують віртуальні ресурси обчислень та зберігання, необхідні для хостингу VNFC, тоді як останній визначає вимоги до продуктивності з'єднання VNFC (VNFC connectivity).

З точки зору атрибутів, VNFD включає в себе один або декілька наступних елементів:

Індикатори VNF: представляють події, пов'язані з продуктивністю/помилками, які надають інформацію про VNF на рівні прикладних програм.

Блоки розгортання віртуалізації (Virtualization Deployment Units VDU): описують як створювати та управляти екземплярами VNFC; отже, VDU можна розглядати як дескриптор VNFC. VDU визначає обчислювальні ресурси (і, можливо, ресурси зберігання), які необхідні контейнеру віртуалізації для розміщення VNFC. З цією метою він посилається на один VCD (і, можливо, один або більше VSD).

Особливості розгортання: подібні до тих, що визначені в NSD, але застосовні до VNFs.

Щоб автоматизувати запуск масштабування, NFVO має настроюваний програмний модуль (наприклад, що підтримує NS-специфічний код), який запускає алгоритм динамічного керування ресурсами (dynamic resource provisioning algorithm DRPA).

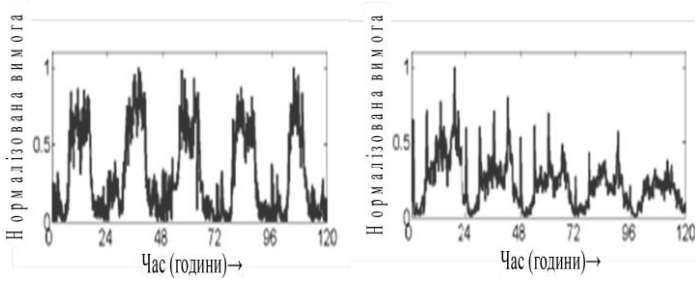
DRPA застосовує відповідні критерії оптимізації (наприклад, мінімізація витрат ресурсів, споживання енергії) та сукупність обмежень (наприклад, доступна ємність ресурсів, обмеження місця розташування) до вихідних даних [51].

2.7 Особливості навантаження та роботи дата центра

Більшість існуючих досліджень проблеми вбудовування віртуальної мережі (Virtual Network Embedding – VNE) в літературі зосереджена на керуванні ресурсами шляхом резервування максимальні потреби в ресурсах для кожної віртуальної мережі протягом всього її життєвого циклу [55], [56].

Визначення ресурсів для сервісів у значній мірі залежить від точної оцінки характеристик обслуговування навантаження. Правильне керування ресурсами є складним завданням через коливання робочого навантаження.

Тим не менш, багато потоків заявок зазвичай мають періодичні шаблони (тобто, щодня, щотижня і/або сезонні цикли). Рис. 2.23 представляє графічно часові ряди запитів протягом п'яти днів, зібраних з програми для підприємства SAP і веб-прикладної програми, відповідно [57].



(а) (б)

Рис. 2.23 Часові ряди для трафіку (а) SAP; (б) Web

Статичний метод керування ресурсами призводить до того, що до 85 відсотків хмарних ресурсів недовикористовуються більшість часу, що призводить до величезних втрат ресурсів і економічних втрат [58]. Завдяки використанню властивості періодичності, коефіцієнт використання ресурсів в хмарі може бути істотно поліпшено.

Проілюструємо це спостереження на прикладі, як показано на рис. 2.24, де Віртуальна мережа A надає офісним користувачам сервіси віртуальних настільних комп'ютерів, і яка як правило, відчуває низькі навантаження у вихідні дні, в той час як інша Віртуальна мережі B надає сервіси онлайн ігор і має високе навантаження в вихідні дні через високі вимоги користувачів. Якщо виділяти ресурси для A і B по їх максимальним вимогам, то тільки одна з них може бути відображена на фізичну мережу. Тим не менш, вони можуть бути обслужені, якщо їх змінні в часі потреби в ресурсах доповнюють один одного.

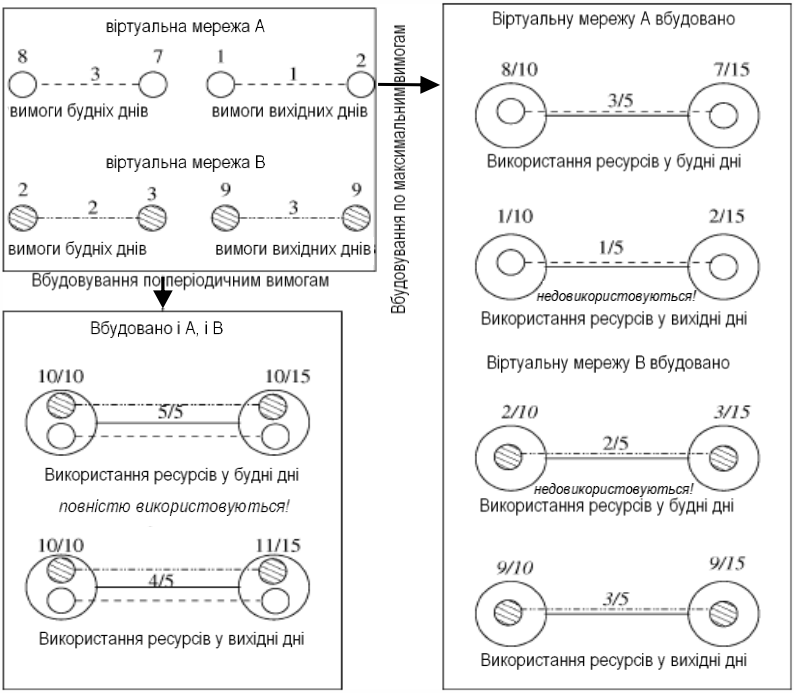


Рис. 2.24 Приклад вбудовування мереж

Якщо визначити шаблони в навантаженні, тоді зможемо внести зміни відповідно до них в розподіл ресурсів і, отже, підвищити точність надання ресурсів і знизити енергоспоживання. Тим не менш, моделі вимог є статистичними по своїй природі, і будуть мати місце відхилення від історичних моделей через непередбачені фактори, такі як несподівані натовпи, перебої в обслуговуванні та святкові дні. Хоча об'єм таких коливань малий в порівнянні з загальною вимогою, повне їх ігнорування може призвести до значних порушень угоди про якість послуг, що надаються, (Service Level Agreement – SLA) [57]. Через мінливість навантаження, що зазнають сучасні системи, розміщення віртуальних машин повинно постійно в режимі реального часу оптимізуватись [59].

Зрештою, надання ресурсів не є безкоштовним [57]; існують різні пов'язані з цим витрати і ризики. Часте виконання процедури надання ресурсів викликає як втрати продуктивності, так і енергії. Наприклад, ввімкнення серверу може зайняти значну кількість часу (до декількох хвилин) і споживати багато енергії (біля пікового споживання) [60]. Часті періодичні вмикання та вимикання потужності серверів викликає «знос», що може призвести до відмови сервера і перебоїв в обслуговуванні [61].

Є інші важливі проблеми, які виникають через високу потужність і енергоспоживання обчислювальних ресурсів. Потужність потрібна для забезпечення роботи охолоджуючої системи. На кожен Ват потужності, споживаний обчислювальними ресурсами, потрібно додатково 0.5-1 Ват для системи охолодження [59]. Крім того, високе споживання енергії інфраструктурою призводить до суттєвих викидів діоксиду вуглецю (CO2), що сприяють виникненню парникового ефекту.

Якщо потреби в ресурсах програми не задовільняються, програма може зіткнутися з підвищеним часом відповіді, тайм-аутами або відмовами. Забезпечення надійної якості обслуговування (QoS), визначеної за допомогою SLA має важливе значення для хмарних обчислювальних середовищ; тому, провайдери мають справу з компромісом по енергоефективності – мінімізацією споживання енергії з одночасним задоволенням SLA [59].

Оскільки навантаження на датацентри коливається з плином часу, можна було б вибірково вимикати частину системи для економії енергії, коли вимоги на систему низькі. Енергозбереження досягається в результаті не тільки від переводу машин в стан сну, але і за рахунок економії витрат на охолодження [62].

Таким чином, навантаження, як правило, має значну мінливість. Це ускладнює надання ресурсів належним чином. Один обсяг (статичне виділення ресурсів) не може підійти всім випадкам, і призведе або до надмірного виділення ресурсів або до недостатнього виділення ресурсів [57].

2.8 Огляд досліджень віртуалізації мережі

Хоча NFV обіцяє істотну економію коштів, гнучкість і простоту розгортання, потенційні проблеми в реалізації віртуалізованих мережевих елементів, які можуть підтримувати вимоги до продуктивності реального світу, і досі залишаються відкритим питанням [63], й в даний час NFV все ще перебуває на початкових етапах реалізації [64]. Предмет віртуалізованого EPC, що може динамічно реконфігуруватись, є досить новим сам по собі. Саме тому в цій сфері не так багато напрацювань [9].

Розподіл ресурсів в NFV мережах подібний до розміщення прикладних програм в датацентрах і хмарах [65]. Задача розміщення функцій також тісно пов'язана з вкладенням віртуальної мережі (VNE) [66]. В останні роки багато наукових досліджень [67] розглядали проблему VNE або застосовуючи математичні моделі оптимізації (наприклад, [8], [22], [68]) або алгоритмічні підходи (наприклад, [69]).

Наприклад, проблема розміщення контролерів розглядалась в [70] і [71], але з основним акцентом на досягнення мінімальної затримки SDN управління, а також стійкість надання послуг. У дослідженні [8] розглядається оптимальне розміщення дата центрів, на яких розміщуються віртуалізовані шлюзи, а також вирішується задача застосування віртуалізації та SDN декомпозиції на шлюзах мобільної базової мережі. Однак задача розміщення виникає і для інших функцій мережі, зокрема EPC. Подібно до проблеми розміщення контролерів в області SDN, такі технології як NFV вимагають відповідних алгоритмів, які можуть вирішувати проблеми, що виходять за рамки однокритеріальних проблем розміщення. Ці проблеми можуть вносити додаткові складності через можливі взаємозалежності між мережевими функціями, як у випадку

об'єднання функцій у ланцюжки, і нові потенційні обмеження відносно додаткових аспектів, таких як безпека [72].

Сучасні дослідження в основному розглядають тільки статичні запити віртуальної мережі, де запит і вимоги на ресурси є фіксованими і не змінюються з часом. Тим не менше, більшість запитів мають динамічні характеристики.

Далі ми розширюємо підхід VNE, визначаючи модель розташування мережевих функцій, що включає в себе поняття гетерогенних мереж, які містять як фізичні засоби обслуговування, так і віртуалізовані сервісні сутності. Схожим чином [22] також фокусується на розгортанні віртуальних мережевих функцій в гетерогенному середовищі, проте не враховує той факт, що продуктивність сервісної сутності залежить від виділених їй ресурсів. Також розглядається питання про те, як оптимально реконфігурувати розгорнуту віртуальну мережу в умовах зміни навантаження.

Також важливо відзначити, що на відміну від більш загальних стратегій керування конфігурацією ресурсів в хмарних середовищах, EPC являє собою граф взаємозалежних вузлів, які не можуть розглядатися ізольовано [9].

Кілька дослідницьких праць були зосереджені на розробці адаптивних систем, які можуть реагувати на зміни у навантаженні в контексті систем зберігання, загальних операційних систем, мережевих сервісів, веб-серверів і Інтернет дата центрів [73]. У даному навчальному посібнику ми розглядаємо абстрактну модель серверного ресурсу і представляємо методи динамічного розподілу ресурсів; запропоновані методи розподілу ресурсів застосовні до багатьох сценаріїв, де система або ресурс можуть абстрагуватись за допомогою GPS (Generalized Processor Sharing) сервера.

Одним з ключових аспектів у області віртуалізації мережі є виділення фізичних ресурсів віртуальним функціям мережі. Це передбачає відображення віртуальних мереж на фізичні мережі, а також керування виділеними ресурсами протягом усього життєвого циклу віртуальної мережі. Ефективність, оптимальність і гнучкість розподілу ресурсів є основними факторами для успішної віртуалізації мережі.

Вбудовування Віртуальної Мережі (VNE) є добре вивченою задачею [66]. Тим не менше, більшість сучасних рішень (наприклад, [74, 75]) пропонують статичну схему розподілу ресурсів, в якій коли віртуальна мережа відображається, перерозподіл ресурсів не відбувається протягом всього її життєвого циклу. Існує обмежена кількість децентралізованих і динамічних рішень VNE (як [76], що розглядає зміни в фізичній мережі, а не зміни в фактичному завантаженні віртуальних мереж, або [77], що пропонує алгоритми для задачі ефективної реконфігурації і вбудовування запитів віртуальної мережі, направлених у хмарний дата центр, автори вимагають, щоб провайдери направляли нові запити на зміну існуючих, і що тільки один такий запит може оброблятися в заданий момент часу). І навіть підходи, які пропонують рішення по динамічному вбудовуванню віртуальної мережі, все одно виділяють фіксовану кількість ресурсів для віртуальних вузлів і каналів на увесь період існування. Оскільки мережевий трафік не є статичним, це може привести до неналежного використання загальних мережевих ресурсів, особливо якщо фізична мережа відхиляє запити на вбудовування нових віртуальних мережевих функцій, при цьому резервуючи ресурси для віртуальних мережевих функцій, які знаходяться в умовах низької завантаженості [78].

Наше дослідження відрізняється від попередніх тим, що перерозподіл ресурсів є проактивним (не запускається невдалими вбудовуваннями), автономним (не запускається

користувачами або мережевими провайдерами) і не передбачає повторні вбудовування вже відображених запитів.

Більшість існуючих робіт по динамічному керуванню ресурсами засновані на трьох підходах: теорії управління, моделюванні динаміки роботи і прогнозуванні навантаження [78]. Серед адаптивних систем, що використовують метод на основі теорії управління [73] і [79]. В системах, що застосовують даний підхід, як правило використовують попередньо визначену модель системи. На відміну від цього, наша методика заснована на онлайн характеризації і прогнозуванні навантаження. Крім того, ці методи використовують залежність між параметром QoS (таким як цільова затримка) і параметром управління (таким як об'єм ресурсів), що не змінюється з часом. Серед робіт, що засновані на динаміці роботи – [80]. Автори [81] використовують прогнозування навантаження.

Підводячи підсумок, різниця між запропонованим у даному дослідженні підходом і згаданими вище є те, що в ньому ресурси, зарезервовані для використання віртуальними мережевими функціями не залишаються незмінними протягом усього життя віртуальної мережі. Здійснюється моніторинг віртуальних вузлів, і на основі їх реальних потреб у ресурсах, ресурси перерозподіляються, і в цьому випадку невикористані ресурси повертаються до фізичної мережі для використання іншими віртуальними мережами.

Наступною задачею, яка виникає, є те, яким чином отримується інформація про поточну ситуацію в мережі. В цьому аспекті керування конфігурацією ресурсів в NFV мережах подібне до керування прикладними програмами в дата центрах і хмарах.

Існуючі рішення керування ресурсами серверів можуть бути класифіковані як прогностичні і реактивні рішення. Прогностичне виділення ресурсів передбачає наявність передбачуваного і стабільного шаблону у вимогах і розподіляє об'єми, як правило, в масштабі часу декількох годин або днів на основі шаблону. У [82] використовується передбачуваність у вимогах бізнес прикладних програм, щоб підвищити ефективність керування ресурсами. Хоча ці підходи можуть бути ефективними в деякій мірі, обирати правильний розмір інтервалу керування досі є дуже складним завданням. Запропонована нами методика дискретизації навантаження може допомогти з цим. Проте, великі, непередбачувані сплески вимог можуть викликати серйозні порушення SLA.

Реактивне виділення ресурсів, з іншого боку, виділяє ресурси в короткі проміжки часу (наприклад, кожні кілька хвилин) у відповідь на зміни навантаження (наприклад, [83]). Чисто реактивні політики потенційно можуть швидко реагувати на зміни навантаження, але такі проблеми, як непередбачуваність, нестабільність і високі витрати керування обмежують їх застосування на практиці [57].

Отже, прогнозувати пік навантаження прикладної програми та виділяти ресурси на основі оцінок найгіршого випадку вкрай складно [84]. З огляду на труднощі в прогнозуванні пікових навантажень, прикладна програма має використовувати комбінацію прогностичного і реактивного керування. У той час як прогностичні методи добре працюють для онлайн прогнозування на великих часових інтервалах до декількох годин, реактивні методи дозволяють прогнозувати навантаження на короткі часові інтервали до декількох хвилин і швидко реагувати на нестаціонарні перевантаження [73].

Існує кілька підходів, які поєднують в собі прогностичне і реактивне керування [57], [84]. Хоча ці підходи мають спільні риси з нашим підходом, вони розрізняються за

кількома аспектами. По-перше, наш підхід спрямований на оптимізацію продуктивності і вартості виділення ресурсів одночасно. По-друге, пропонується метод аналізу та дискретизації навантаження, для визначення оптимального розміру інтервалів часу сталої конфігурації ресурсів – інтервали мають змінну довжину, тоді як в інших підходах до керування використовуються прості фіксовані інтервали [85]. [57] схожим чином використовує змінну довжину інтервалів, однак на противагу цьому підходу, запропонований модифікований підхід довжину інтервалів визначає динамічно в залежності від фактичної ситуації в мережі.

Однією з основних проблем в динамічній мережі з NFV є те, як здійснити розумне розгортання VNF, щоб адаптуватися до змін в мережі. Маневрений механізм розгортання мережевих функцій може бути використаний для вирішення цієї проблеми. Коли змінюється мережа, адміністратор може видалити застарілі VNF і перерозподілити нові в кращому місці.

Існуючі роботи, пов'язані з розгортанням VNF і міграцією, як правило, зосереджені на пропозиціях нових стратегій розгортання [86] і механізмів міграції [87] [88]. Але вартість міграції не розглядається в цих дослідженнях. Насправді, вартість міграції є ключовим фактором в процесі міграції. Вона буде впливати на рішення про те, як вибрати кандидат VNF для перенесення і з'ясувати цільову позицію, яка підходить для міграції [89].

Стійкість обслуговування є важливою вимогою в будь-якій системі зв'язку, особливо в мобільних мережах, відомих своєю надійністю п'ять дев'яток. Крім того, доступність і надійність послуг, як зазначено в Рекомендації МСЕ-Т Е.800 [90], визначають ключовий параметр характеристик якості обслуговування (QoS). В контексті хмарних мереж зв’язку, забезпечувати стійкість сервісів стає проблемою. Дійсно, висока доступність є важливою вимогою, але не обов'язково є невід'ємною рисою хмарних обчислень. Побудова системи, яка вимагає надійності п'яти дев'яток на платформі, яка не завжди може надати її, є таким чином перешкодою. В хмарному середовищі на стійкість обслуговування може в значній мірі впливати відмова будь-якої VNF, що працює на віртуальній машині. Відмова VNF може статися через кілька факторів, таких як відмова апаратних засобів (наприклад, через неправильне масштабування обладнання), вразливості програмного забезпечення та помилки в управлінні VNF (тобто, в основному, якщо VNF складається з декількох компонентів, кожен з яких розгорнутий на своїй власній віртуальній машині на тому ж або іншому апаратному забезпеченні) або її відповідній віртуальній машині, відмова на рівні гіпервізора через неправильну конфігурацію, негативний вплив на продуктивність за рахунок інших VNF розміщених на одному фізичному вузлі, і злонавмисних атак проти VNF або менеджера віртуальної машини (тобто, гіпервізора) [91]. В операторській хмарі, відмова VNF може вплинути на площину управління (наприклад, MME), а також на площину даних користувача (наприклад, S-GW або P-GW). У площині управління, роль MME має вирішальне значення, так як він відповідає за численні важливі процедури (наприклад, встановлення з'єднання з великим числом вузлів/VNF площини даних користувача, обладнання користувача (User Equipment – UE) – управління мобільністю і аутентифікацію UE). Його відмова істотно впливає на надання послуг, таким чином, важливість вивчення стійкості сервісів EPCaaS шляхом визначення оперативних, масштабованих і надійних механізмів відновлення для відновлення після збоїв VNF [92].

Одна з основних проблем в контексті віртуалізації мережі полягає в тому, як ефективно використовувати нижчерозташовані фізичні ресурси (тобто центральний процесор і пам'ять вузлів і смугу пропускання каналів). Відображення віртуальних вузлів і віртуальних каналів на фізичні ресурси, як відомо, є NP-важкою задачею [93].

Кілька дослідницьких спроб [94] вирішення цієї задачі були представлені; ряд цих досліджень вводив різні методи для вирішення завдання відображення віртуальної мережі в надії встановлення ефективного використання фізичних ресурсів. У доповненні до методів, необхідних для ефективного відображення віртуальних мереж на фізичні мережі, потрібні методи, які керують ресурсами, вже виділеними активним віртуальним мережам. Отже, існує необхідність розробити методику, яка може перемістити вже розміщені віртуальні вузли в разі відмови вузла або обслуговувасння вузла при зведенні до мінімуму періоду переривання обслуговування.

У той час як адаптація шляхів для віртуальних мереж була розглянута у ряді підходів [95], проблема відмови вузла в віртуальних мережах була розглянуто раніше лише у [96], проте не враховувалася вартість ресурсів на вузлі і кінцева якість обслуговування, а також не розглянуто випадок відмови вузла через надмірне навантаження, що надходить на нього. Крім того, невирішеною залишалась задачу вибору місць розташування вузлів керування.

У даному навчальному посібнику, представляється розподілений алгоритм локальної реконфігурації мережі, який гнучко перерозподіляє віртуальні вузли, які постраждали від збою або перевантаження на фізичному вузлі. Основна мета полягає в розробці механізму самовідновлення віртуальної мережі, який може мінімізувати період переривання обслуговування і вартість відновлення вузла після відмови, а також підтримувати високий рівень фізичної працездатності мережі, що, в свою чергу, збільшує прибуток провайдера.

Таким чином, на відміну від існуючої статичної архітектури мережі LTE EPC пропонується система (рис. 2.25), в якій виділені апаратні мережеві функції обслуговують певний заданий рівень службового навантаження (штрихова лінія на рис. 2.26), в той час як заявки службових потоків, що цей рівень перевищують (крива на рис. 2.26), направляються на обробку з використанням віртуалізованих мережевих функцій у орендованих хмарах дата центрів.

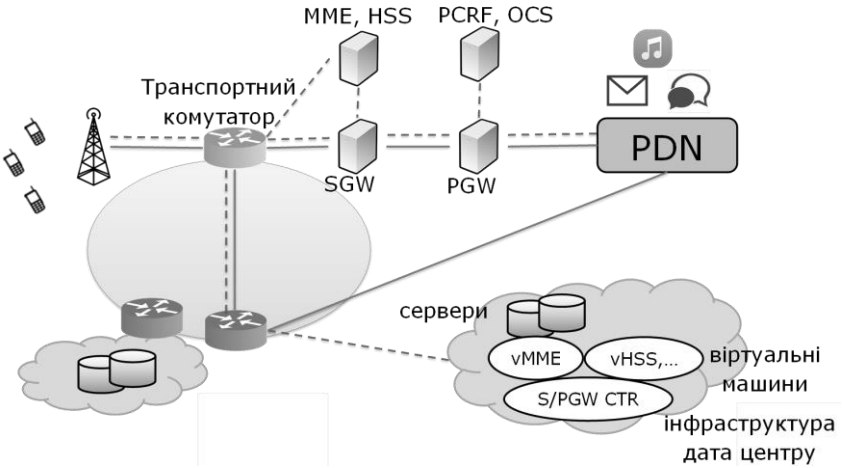


Рис. 2.25 Архітектура мережі LTE EPC з використанням NFV (змінна залежно від обсягів навантаження)

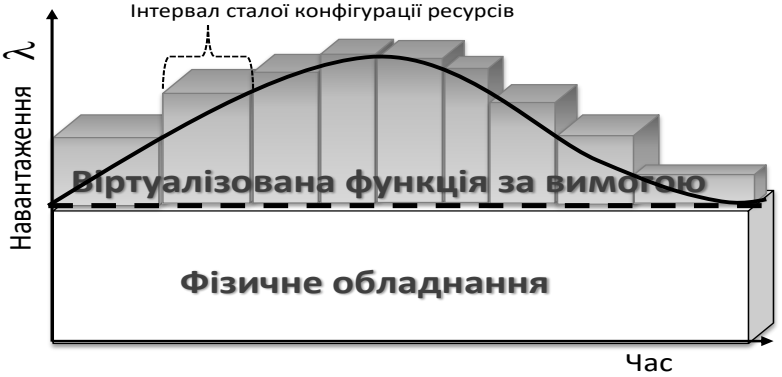


Рис. 2.26 Розподіл навантаження в гетерогенній мережі

Далі описано задачу визначення місця розміщення та необхідного обсягу віртуальних зарезервованих ресурсів у контексті віртуалізації мережевих функцій. В наступних розділах описано підхід до керування інтервалом сталої конфігурації ресурсів з урахуванням прогнозу навантаження і пропонований метод для локальної реконфігурації ресурсів віртуалізованих мережевих функцій.

Висновки

1. В результаті розгляду основних тенденцій технологій мобільних мереж майбутнього виявлено, що мобільні системи зв'язку наступних поколінь будуть застосовувати середовище з високими вимогами до якості зв'язку, тому для надання послуг і плавного переходу від одного покоління мобільних мереж до іншого, розроблена концепція реконфігурованості.
2. Визначення ключових характеристик застосування системи динамічного керування конфігурацією ресурсів показало доцільність її використання для вирішення актуальних телекомунікаційних проблем, хоча такі фактори, як підвищена складність роботи і проблеми регулювання в порівнянні з перевагами даного підходу повинні бути ретельно проаналізовані, перш ніж застосовувати його у системі мобільного зв'язку.
3. Проведений огляд сучасного телекомунікаційного середовища показав, що тенденції, такі як нестаріюче обладнання, безшовна інтеграція нових сервісів та багаторежимне обладнання у галузі комерційних безпроводових мереж привели до широкого інтересу до технології NFV.
4. Визначено, що для задоволення ряду суперечливих вимог до апаратних елементів необхідно застосовувати гетерогенні системи, які дозволяють спільно використовувати різні технології, що забезпечує найефективніше виконання кожної функції.

Контрольні запитання

1) Які основні відмінності в способах реалізації надання віртуалізованого мережевого сервісу в порівнянні з існуючою практикою?
2) Які існують спонукальні чинники зростання сигнального трафіку?
3) Охарактеризуйте економічні вигоди від віртуалізації мобільної мережі.
4) З яких вузлів складається архітектура EPC?
5) Поясніть типову архітектуру LTE/EPC.
6) Які функції виконує вузол MME?
7) Які функції виконує вузол HSS?
8) Які функції виконує вузол SGW?
9) Які функції виконує вузол PGW?
10) Які функції виконує вузол PCRF?
11) Які є основні категорії QoS?
12) Що таке ідентифікатор класу якості обслуговування (QCI) і пріоритет розподілу і утримання (ARP)?
13) З чого складається високорівнева платформа NFV?
14) Що таке MANO?
15) Які функції виконує VIM?
16) Які функції виконує VNFM?
17) Які функції виконую NFVO?
18) Що таке мережева функція?
19) Що таке мережевий сервіс?

20) Що таке хмарні обчислення?
21) Які ключові характеристики хмарних обчислень?
22) Які є основні моделі хмарних обчислень?
23) Який взаємозв'язок між хмарними обчисленнями та NFV?
24) Який взаємозв'язок між NFV, SDN та хмарними обчисленнями?
25) Що таке дескриптор мережевого сервісу?
26) Яку структуру має NSD?
27) Які особливості навантаження та роботи дата центра?
28) Як існуючі рішення керування ресурсами серверів можуть бути класифіковані
?
29) Як можна побудувати мережу LTE EPC з використанням NFV?

РОЗДІЛ 3
РЕЗЕРВУВАННЯ РЕСУРСІВ ДЛЯ ВІРТУАЛІЗОВАНИХ МЕРЕЖЕВИХ ФУНКЦІЙ В ГЕТЕРОГЕННОМУ СЕРЕДОВИЩІ

3.1 Процедури LTE та потік заявок

Абонентське обладнання UE – це термінали, які дозволяють кожному користувачеві підключатися до мережі через базові станції eNodeB. На UE виконуються прикладні програми користувачів, які генерують або споживають мережевий трафік. UE може ініціювати запити до мережі, використовуючи керуючі повідомлення. Діяльність UE та генерація трафіку мережі також викликають процедури керування мережею. eNodeB отримують повідомлення UE до EPC.

Для забезпечення функціонування стандарт LTE визначає процедури сигналізації, які передбачають обмін сигнальними повідомленнями між об'єктами LTE (наприклад, eNB, MME, S-GW та HSS).

Для опису та розуміння того, як працює телекомунікаційна система, стандартизовано процедури мережі LTE. Детальний опис процедур викладено у технічних специфікаціях 3GPP TS 23.060 [97], 23.401 [98], і 23.402 [99]. Коротко розглянемо, для прикладу, процедуру підключення UE [100].

Підключення є першою процедурою, яку UE виконує після включення. Процедура виконується, щоб зробити можливим отримання послуг від мережі. Оптимізацією в системі SAE є те, що процедура підключення також включає в себе встановлення каналу EPS за замовчуванням, гарантуючи, що завжди на зв'язку IP зєднання для UE/користувачів EPS доступне. Приклад процедури підключення описаний на рис. 3.1.

Процедура коротко описана в наступних кроках:

A. UE відправляє повідомлення запиту Attach Request до eNodeB. eNodeB перевіряє ідентифікатор MME переданий на рівні the Radio Resource Control (RRC). Якщо eNodeB має канал до визначеного MME, він направляє запит Attach Request на цей MME. Якщо не має, то eNodeB вибирає новий MME і пересилає запит Attach Request.

B. MME змінився і MME використовує старий MME ID в GUTI, щоб знайти старий MME і отримує контекст UE.

C. Виконуються процедури аутентифікації та забезпечення безпеки. Визначення ME також виконується в поєднанні з цим кроком.

D. Якщо MME змінився, MME повідомляє HSS, що UE перемістився. HSS зберігає адресу MME і інструктує старий MME скасувати контекст UE.

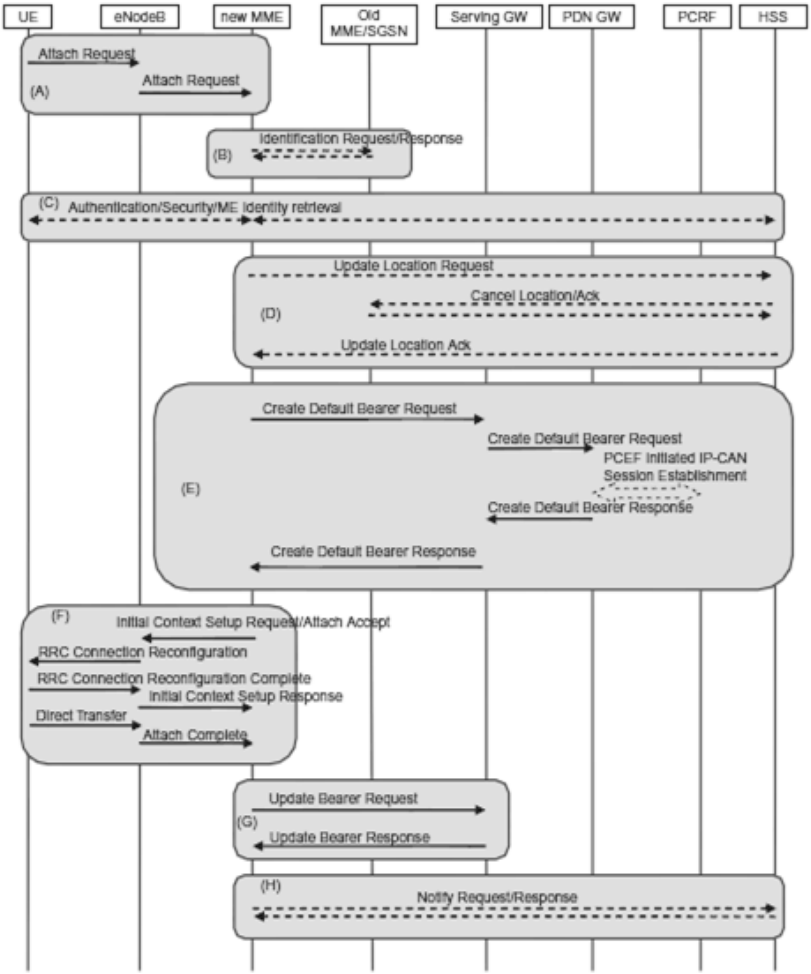


Рис. 3.1 Процедура підключення

E. Канал за замовчуванням авторизується PCRF і встановлюється між SGW і PGW.

F. Канал за замовчуванням встановлюється через радіоінтерфейс і підтвердження Attach Accept надсилається в UE.

G. MMA інформує SGW про ідентифікатор кінцевої точки тунелю (TEID) eNodeB, який завершує установку каналу за замовчуванням, так що він тепер може бути використаний як у висхідній лінії зв'язку, так і у низхідній лінії зв'язку.

H. Якщо MME вибрав PGW, який не є таким же, як той, що вказано в отриманій інформації підписки, то він надішле повідомлення про новий PGW на HSS.

Крім того, існують деякі додаткові кроки, які можуть бути виконані разом з процедурою підключення. Наприклад, якщо тимчасовий ідентифікатор UE (GUTI) невідомо як в старому MME, так і в новому MME (після кроків A і B), новий MME проситиме UE відправити свій постійний ідентифікатор підписки (IMSI), як показано на рис. 3.2.

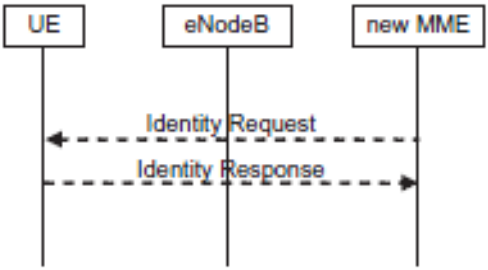


Рис. 3.2 Запит ідентифікатору

MME може перевірити справжність ME за допомогою регістру ідентифікації обладнання (EIR; після кроку C). EIR може бути використаний для чорного списку, наприклад, вкрадених UE. Залежно від відповіді від EIR, MME може продовжити процедуру підключення або відхилити UE (див рис. 3.3).

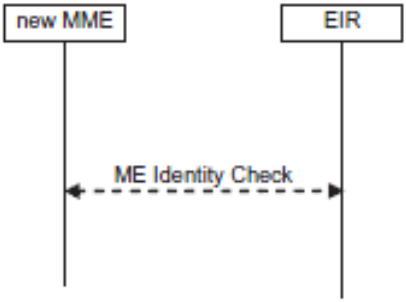


Рис. 3.3 Перевірка ідентифікатору

Якщо UE захотів відправити APN або PCO, він встановлює прапор в початковому повідомленні запиту Attach Request. MME потім запитує інформацію від UE після

початку шифрування на етапі С. Таким чином, немає ніякої необхідності відправляти APN або PCO у незашифрованому вигляді через інтерфейс радіозв'язку. Варіанти шифрування процедури запиту використовуються для передачі APN і/або PCO в MME (див. рис. 3.4).

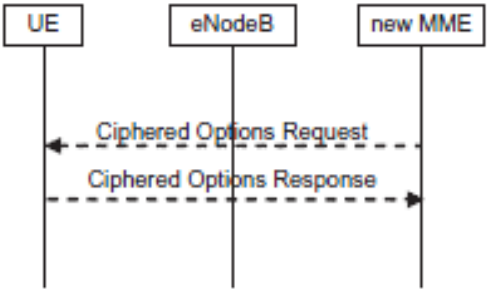


Рис. 3.4 Запит опцій шифрування

Потік заявок – це послідовність заявок, що надходять на обслуговування в систему в деякі моменти часу. Таким чином, кожен з типів запитів процедур LTE формує пуасонівський потік заявок з деякими інтенсивностями ν_1, ν_2, ... ν_{proc}. Маємо суму кінцевого числа *proc* пуасонівських процесів, що формує сумарний пуасонівський потік з параметром інтенсивності $\lambda=\nu_1+\nu_2...+\nu_{proc}$

Середню інтенсивність вхідного потоку заявок λ до системи для первинного розміщення мережевих функціональних блоків визначаємо виходячи з відомої середньої активності одного абонента (Λ):

$$\lambda=\Lambda \cdot Quantity,$$

де кількість абонентів *Quantity* на території, що охоплюється, розраховується виходячи зі щільності населення ρ, що користується послугами обраного оператора: $Quantity=\rho \cdot S$, S – площа території.

Як приклад мережевої функції розглянемо блок MME. MME є основним елементом керування мережею. Він відповідає за збереження стану мобільності UE, управління каналом, аутентифікацію та авторизацію користувачів та інші функції. Коли MME одержує одне сигнальне повідомлення, він обробляє його, а пізніше надсилає нове повідомлення іншому об'єкту (такому як eNB або S-GW). Якщо процедура вимагає кількох етапів, сутність надсилає інше повідомлення до MME.

Існує кілька процедур сигналізації в LTE. З усіх них розглянемо ті, які генерують більшість сигнальних навантажень [101].

1) Service Request (SR): коли UE не має доступних ресурсів і створюється новий трафік, або від цього UE або від мережі до цього UE, UE виконує процедуру запиту сервісу Service Request (SR). Розглянемо SR, що запускається UE. Під час цієї процедури MME отримує три різні повідомлення (рис. 2.5): Initial UE Message (SR_1), Initial Context Setup Response (SR_2), Modify Bearer Response (SR_3).

Для обробки Initial UE Message (SR_1) MME повинен виконувати перевірку цілісності UE та дешифрування повідомлення. Крім того, він генерує ідентифікатори для каналів, які мають бути встановлені. Крім того, він зберігає та надає параметри та змінні, пов'язані з контекстом UE. Деякі з них включаються в наступне повідомлення Initial Context Setup Request. Під час обробки повідомлення Initial Context Setup Response (SR_2), MME також отримує інформацію про контекст UE та включає цю інформацію в наступне повідомлення Modify Bearer Request. Обробка Modify Bearer Response (SR_3) є мінімальною, оскільки це повідомлення є лише підтвердженням.

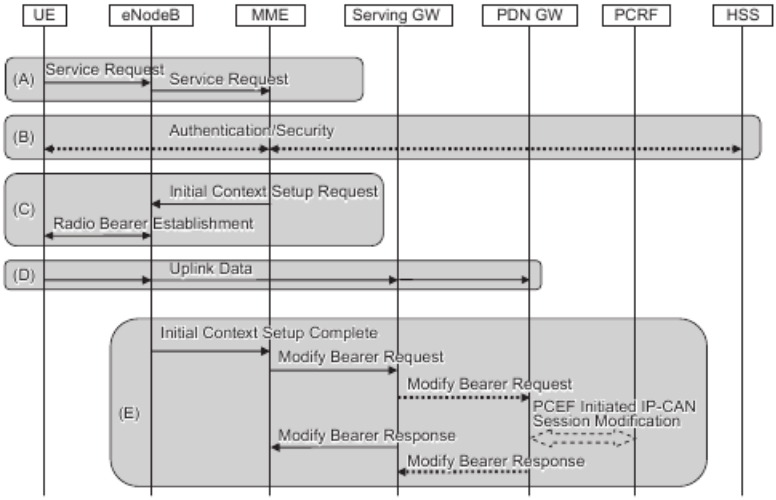


Рис. 3.5 Процедура «Запит сервісу» ініційована UE

2) Service Release (SRR): процедура Service Release (SRR) запускається неактивністю користувача. Її мета полягає в тому, щоб звільнити радіоканал даних та низхідний S1 канал у площині даних, а також сигнальні з'єднання радіо та S1 у площині управління для UE. Під час SRR MME обробляє три повідомлення (рис. 3.6): UE Context Release Request (SRR_1), Release Access Bearers Response (SRR_2), UE Context Release Complete (SRR_3).

Для обробки повідомлень UE Context Release Request message (SRR_1) та Release Access Bearers Request (SRR_2), MME має отримати інформацію про контекст UE та включити цю інформацію в наступні повідомлення. Обробка повідомлення UE Context Release Complete (SRR_3) в основному означає видалення контекстної інформації каналу.

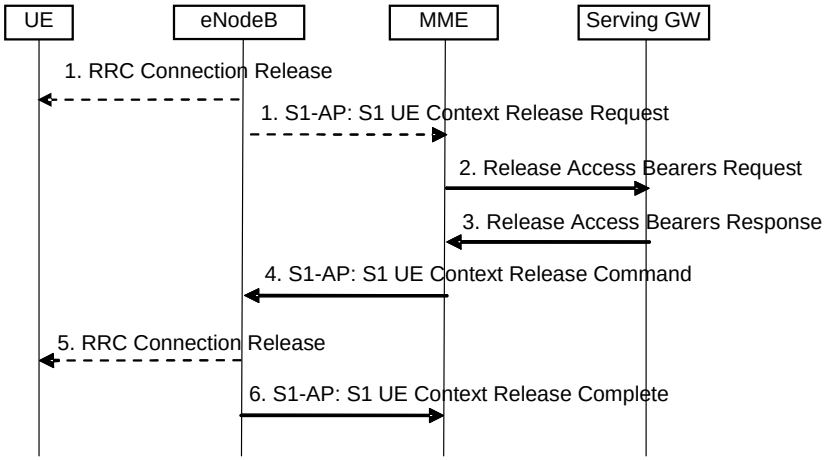


Рис. 3.6 Процедура «S1 звільнення»

3) X2-Based Handover (HR): MME бере участь у X2-Based Handover під час фази завершення хендоверу. Його мета - переключити кінцеву точку каналу від вихідної до цільової eNB. У цьому етапі MME отримує два повідомлення: Path Switch Request (HR_1) та Modify Bearer Response (HR_2) (рис. 2.7).

Для обробки цих повідомлень MME також повинен отримати інформацію про контекст UE і включити цю інформацію в наступні повідомлення. Для обробки повідомлення Path Switch Request, MME також необхідно зберігати нову інформацію, таку як ідентифікатори нового обслуговуючого стільника та нову область відстеження.

Нехай λ – сумарна інтенсивність надходження повідомлень сигналізації для блоку MME. Тоді, відповідно до опису процедур управління, λ розраховується як

$\lambda=\nu_{SR}+\nu_{SRR}+\nu_{HR}=3\lambda_{SR}+3\lambda_{SRR}+2\lambda_{HR}$,

де λ_{SR}, λ_{SRR} та λ_{HR} позначають середні значення інтенсивності надходження для процедур SR, SRR та HR, відповідно. Ці показники можна виразити в термінах середніх значень інтенсиностей для користувача Λ^h_U для процедури $h \in \{SR, SRR, HR\}$. Отримаємо $\lambda_h = \Lambda^h_U \cdot Quantity$.

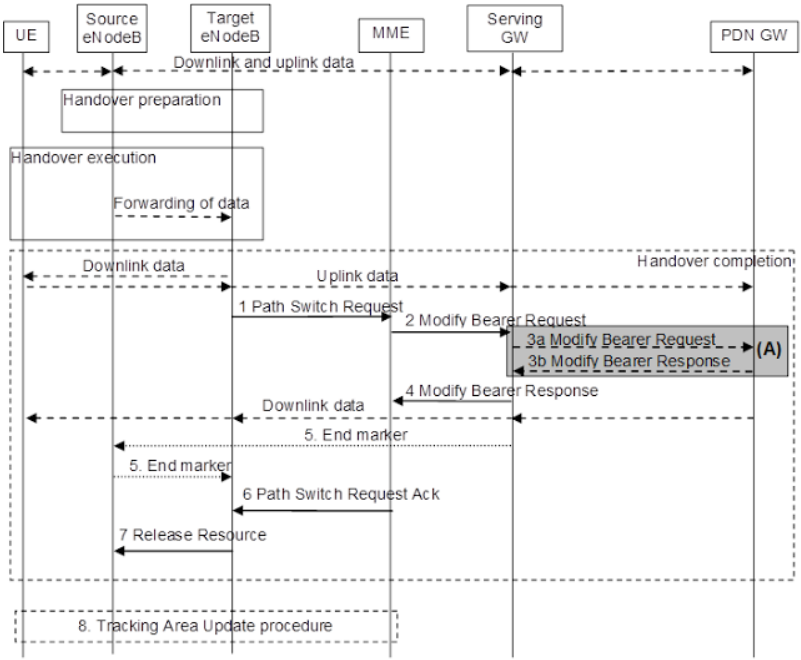


Рис. 3.7 Процедура «X2 хендовер без переміщення SGW»

3.2 Визначення кількості сервісів віртуальних мереж

Кількість сервісних ланцюгів потрібно визначити заздалегідь [67]. Крайнім випадком був б розгляд одного сервісного ланцюга на стільник/eNodeB. Оскільки реалістичні сценарії для мобільних мереж складають 10000 eNodeB, результуюча оптимізаційна модель буде величезна, і для її вирішення потрібен досить тривалий час обчислення. Тому приймаємо обгрунтовано великі кластери eNodeBs і припустимо, що кожен з цих кластерів eNodeB звертається до одного сервісного ланцюга базової мережі.

Більшість сучасних досліджень не розглядає задачу визначення кількості мереж EPC як сервісів, а передбачається їх наперед задана кількість. Найбільш близькою виступає задача з дослідження [102], яка, проте, направлена на визначення розміщення та прив'язки RRH сайтів до BBU серверів у мережах з застосуваннях технології NFV і не враховує обмежень затримки та пропускної здатності.

Далі описано алгоритм вибору вузлів агрегації трафіку. Зокрема, розглядаємо випадок, коли провайдер телекомунікаційних послуг вже має існуючу топологію базових

станцій. Потрібно визначити підмножину мережевих вузлів, де будуть розміщені блоки агрегації навантаження, які будуть формувати запити до одного віртуалізованого сервісу EPC. Після цього для кожного сайту базовихстанцій призначаємо вузол агрегації (Traffic Aggregation Point – TAP).

3.2.1 Постановка задачі вибору вузлів агрегації трафіку

Задача проілюстрована на рис. 3.8. На цьому рисунку певна кількість базових станцій групується та призначається до одного TAP.

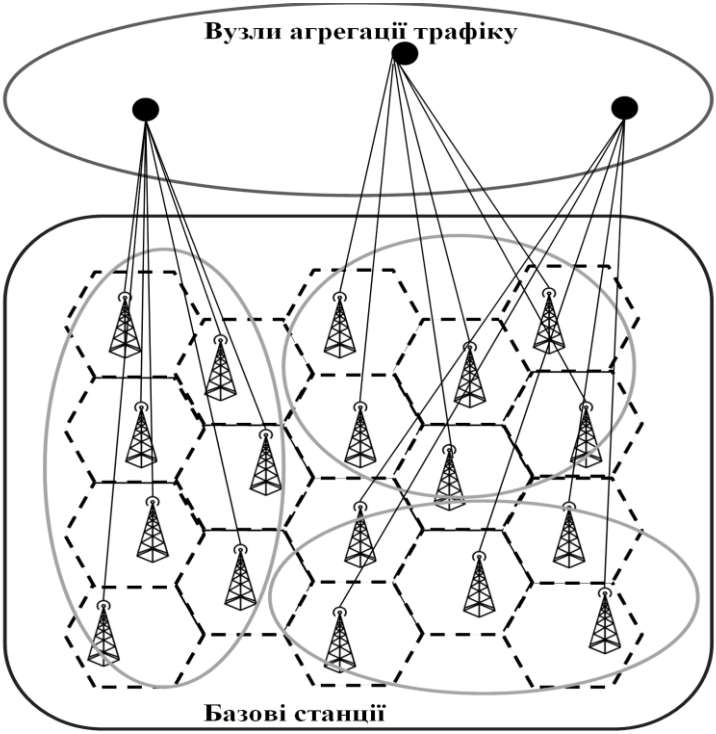


Рис. 3.8 Представлення проблеми розміщення вузлів агрегації

Сайти eNodeB

Розглядаємо мережу з множиною J сайтів eNodeB, які повинні обслуговуватись іншою множиною I вузлів агрегації, де $|I| \leqslant |J|$. Розташування кожного сайту $j \in J$ характеризується географічною широтою та довготою. Кожен сайт базової станції $j \in J$

надає послуги абонентам, чиї комбіновані середні вимоги до обробки є d_j. Для кожного сайту визначено максимальний рівень затримки T_j, який задає міру максимальної відстані від сайту до ТАР.

ТАР

Вважаємо, що провайдер може встановити максимум $p = |I| \leqslant |J|$ фізичних ТАР для обслуговування всіх сайтів мережі радіодоступу. Кожен ТАР $i \in I$ має максимальну ємність обробки P_i. Ця максимальна обчислювальна потужність спрямована на те, щоб відобразити той факт, що на практиці неможливо мати ТАР з необмеженим обсягом обчислювальних ресурсів. Розміщення ТАР $i \in I$ включає в себе вартість $cost_i$, яка може бути визначена як лінійна функція вартості, показана у (3.1). Вартість складається з двох частин: фіксованої початкової вартості f_i, яка відповідає за фіксовані інвестиції, такі як простір та встановлення обладнання, а також додаткові витрати $costN_i$ на одиницю обробної потужності, встановленої на обчислювальному вузлі, де d_i – обсяг задіяних обчислювальних ресурсів обробки.

$$cost_i = f_i + costN_i \cdot d_i \tag{3.1}$$

Канали зв'язку

Кожний сайт підключений до відповідного ТАР через лінію зв'язку лінії $e_{ji} \in E$, де E являє собою набір всіх ліній зв'язку. Кожна лінія зв'язку e_{ji} має затримку L_{ji}, яка залежить від відстані між сайтом j і ТАР i та швидкістю сигналів у використовуваному транспортному середовищі, а також доступну пропускну здатність c_{ji}. Вартість встановлення зв'язку між сайтом j і ТАР i дорівнює $costl_{ji}$ і визначається як лінійна комбінація початкової фіксованої вартості fl_{ji} та змінної частини залежної від смуги пропускання B_{ji}, необхідної каналу, та вартості одиниці пропускної здатності $costL_{ji}$, як показано в (3.2).

$$costl_{ji} = fl_{ji} + costL_i \cdot B_{ji} \tag{3.2}$$

3.2.2 Формулювання оптимізаційної задачі

Визначення змінних рішення

Нехай x_i – це бінарна змінна рішення, яка приймає значення 1, якщо в точці i необхідно розмістити ТАР, та 0 – в іншому випадку. Крім того, визначаємо y_{ji} як бінарну змінну, яка приймає значення 1, якщо базова станція j направляє навантаження в i-у ТАР, і 0 – в іншому випадку. Необхідно визначити значення x_i та y_{ji}, так щоб знайти оптимальне значення цільової функції.

Цілі оптимізації

Цільова функція (3.3) прагне мінімізувати мережеві затримки. Цільова функція (3.4) представляє загальну умовну вартість встановлення вузлів агрегації трафіку та вартість встановлення каналів між базовими станціями та відповідними ТАР, що їх обслуговують. Цільова функція (3.5) має на меті залишати більше вільної пропускної здатності на кожному фізичному каналі. Максимізується залишкова пропускна здатність по всіх каналах, оскільки значно завантажені канали можуть призвести до перевантажень у мережі, тому бажано отримати рішення, де залишається більше потужностей каналів.

Ці цілі оптимізації можуть бути корисними для мережевих операторів для планування найкращої стратегії розгортання.

$$\min_{x_i,y_{ji}}\left(\sum_i \sum_j y_{ji} \cdot L_{ji}\right), \quad (3.3)$$

$$\min_{x_i,y_{ji}}\left(\sum_i x_i \cdot cost_i + \sum_i \sum_j y_{ji} \cdot costl_{ji}\right), \quad (3.4)$$

$$\max_{x_i,y_{ji}}\left(\sum_i y_{ji} \cdot (c_{ji} - B_{ji})\right) \quad (3.5)$$

Можливо використовувати лінійну комбінацію (3.6) виразів (3.3)-(3.5) з ваговими коефіцієнтами *a*, *b*, *c*, які можуть використовуватись не лише щоб задати більшу вагомість тої чи іншої складової, але також щоб масштабувати значення виразів з метою зведення до порівнюваних значень та мати значуще складення.

$$\min_{x_i,y_{ji}}\begin{pmatrix} a \cdot \sum_i \sum_j y_{ji} \cdot L_{ji} + b \cdot \left(\sum_i x_i \cdot costN_i + \sum_i \sum_j y_{ji} \cdot costL_{ji}\right) - \\ -c \cdot \left(\sum_i y_{ji} \cdot (c_{ji} - B_{ji})\right) \end{pmatrix} \quad (3.6)$$

Обмеження

1. Розташування та призначення:

Обмеження (3.7) гарантує, що кожна базова станція буде приєднана до одного ТАР. Обмеження (3.8) гарантує, що між сайтом базової станції *j* та ТАР *i* створюється канал, тільки якщо *i* було розміщено. Разом обмеженнями (3.7) та (3.8) забезпечується, що розміщується потрібна кількість ТАР для обслуговування всіх сайтів, а базові станції призначаються лише вузлам, на яких розміщені ТАР.

$$\sum_i y_{ji} = 1 \ \forall j, \quad (3.7)$$

$$y_{ji} \leq x_i \ \forall j \ \forall i \quad (3.8)$$

2. Забезпечення допустимості обсягу ресурсів:

Обмеження (3.9) гарантує, що максимальна кількість ТАР не перевищує бюджет p, тоді як (3.10) є обмеженням потужності, що гарантує, що загальні вимоги до обробки всіх базових станцій, призначених для конкретного ТАР, не перевищують фактичні встановлені фізичні ресурси. Обмеження (3.11) гарантує достатність канальних ресурсів для встановлення каналів, а (3.12) – допустимість значення затримки.

$$\sum_i x_i \leq p, \quad (3.9)$$

$$\sum_j y_{ji} \cdot d_j \leq p_i \ \forall i, \quad (3.10)$$

$$\sum_i y_{ji} \cdot (c_{ji} - B_{ji}) \geq 0 \ \forall j, \quad (3.11)$$

$$\sum_i y_{ji} \cdot L_{ji} \leq T_j \ \forall j \quad (3.12)$$

Вирішити задачу (3.1)-(3.12) можна з застосуванням евристичних методів оптимізації (наприклад, генетичного алгоритму тощо).

3.3 Відображення віртуальних вузлів на фізичні вузли

Підхід базується на спільному розташуванні індивідуальних ланцюгів сервісів базової мережі на фізичній мережі, де ланцюг сервісів позначає мережеві функції мобільної базової мережі, яку потік трафіку повинен пройти [67]. Припускаємо, що віртуальні мережеві функції мають таку ж функціональність і інтерфейси як і мережеві елементи архітектури 3GPP LTE EPC.

Фізична мережа задана у вигляді графа *SN = (N,NE)*, де *N* є множиною фізичних вузлів і *NE* – множиною каналів. Кожен канал $l = (n_1,n_2) \in NE$, $n_1,n_2 \in N$ має максимальну пропускну здатність $c(n_1,n_2)$ і кожен вузол $n \in N$ пов'язаний з певними ресурсами c_n^i, $i \in R$, де *R* – множина типів ресурсів. Множина усіх точок агрегації трафіку (Traffic Aggregation Point – TAP), тобто кластерів eNodeB, в мережі позначається $K \subseteq N$. Для

кожного вузла $n \in N$, $suit_n^{k,j}$ є бінарним параметром, який вказує, чи адміністративно можливо розгорнути на вузлі функцію типу $j \in V$, де V є множиною типів мережевих функцій, k-го сервісу, де $k \in K$.

Віртуальна базова мобільна мережа представляється множиною сервісів (один сервіс на ТАР), які вбудовуються в фізичну мережу.

Вимоги смуги пропускання каналу між двома функціями, *j1* і *j2*, $(j1,j2) \in E$, що відносяться до сервісу ТАР $k \in K$ позначається як $d_k^{(j1,j2)}$. $d_k^{j,i}$ – кількість ресурсу типу i, що виділяється для мережевої функції j сервісу k. $s_{n,i}^{k,j}$ позначає час обробки запиту на ресурсі типу i мережевої функції j сервісу k однією одиницею ресурсу вузла n. Вимоги до допустимого часу обробки заявки мережевою функцією j, що відноситься до сервісу k, позначаються як P_k^j. T_k - максимальна затримка для $k \in K$, $L(n_1,n_2)$ – мережева затримка для каналу $(n_1,n_2) \in NE$.

Метою оптимізації є знаходження розташування віртуалізованих сервісів базової мережі (тобто розміщення мережевих функцій та розподіл ресурсів, а також визначення шляхів передачі трафіку між ними), так щоб мінімізувати витрати на зайняті ресурси каналів і вузлів у фізичній мережі, при цьому задовільняючи вимоги трафіку. Сформулюємо цільову функцію (вираз (3.13)) у вигляді лінійної комбінації двох вартісних виразів: зайнятого обсягу ресурсів обчислювальних вузлів, де умовна вартість одиниці ресурсу i на вузлі n позначається як $costN(i,n)$, і зайнятої пропускної здатності каналів, де $costL(n_1,n_2)$ – умовна вартість одиниці пропускної здатності фізичного каналу $(n_1,n_2) \in NE$.

Наступні формули (3.13)-(3.22) представляють собою постановку оптимізаційної задачі нелінійного програмування. Змінні $x_n^{k,j}$ вказують, чи мережева функція j пов'язана сервісом k розташовується на фізичному вузлі n. Для j=ТАР, $x_n^{k,TAP}$ – не змінні, а вхідні параметри, які вказують де ТАР k знаходиться, тобто

$$x_n^{k,TAP} = \begin{cases} 1 \text{ якщо } k = n, \\ 0 \text{ інакше} \end{cases}.$$

Аналогічно, змінні $f_{(n_1,n_2)}^{k,(j_1,j_2)}$ вказують, чи фізичний канал $(n_1,n_2) \in NE$ використовується для шляху між j_1 і j_2 для сервісу k.

$$\min_{x_n^{k,j}, f_{(n_1,n_2)}^{k,(j_1,j_2)}, d_t^{j,i}} \Big(\sum_{k \in K} \sum_{j \in V} \sum_{n \in N} \sum_{i \in R} x_n^{k,j} \cdot d_k^{j,i} \cdot costN(i,n) + \\ + \sum_{(n_1,n_2) \in NE} costL(n_1,n_2) \cdot \sum_{k \in K} \sum_{(j_1,j_2) \in E} f_{(n_1,n_2)}^{k,(j_1,j_2)} \cdot d_k^{(j_1,j_2)} \Big) \quad (3.13)$$

$$\text{З обмеженнями } \sum_{n \in N} x_n^{k,j} = 1 \ \ \forall k \in K, j \in V, \quad (3.14)$$

$$x_n^{k,j} \le suit_n^{k,j} \ \ \forall k \in K, j \in V, n \in N, \quad (3.15)$$

$$\sum_{(w,n) \in NE} \sum_{k \in K} \sum_{(j_1,j_2) \in E} f_{(w,n)}^{k,(j_1,j_2)} \cdot d_k^{(j_1,j_2)} \le c_n^{bdw} \ \ \forall n \in N, \quad (3.16)$$

$$\sum_{k \in K} \sum_{j \in V} x_n^{k,j} \cdot d_k^{j,i} \le c_n^{i} \ \ \forall n \in N, i \in \{R \backslash bdw\}, \quad (3.17)$$

$$\sum_{k \in K} \sum_{(j_1,j_2) \in E} f_{(n_1,n_2)}^{k,(j_1,j_2)} \cdot d_k^{(j_1,j_2)} \le c(n_1,n_2) \ \ \forall (n_1,n_2) \in NE, \quad (3.18)$$

$$\sum_{(n,w) \in NE} f_{(w,n)}^{k,(j_1,j_2)} - f_{(n,w)}^{k,(j_1,j_2)} = x_n^{k,j_1} - x_n^{k,j_2} \ \ \forall k \in K, n \in N, (j_1,j_2) \in E, \quad (3.19)$$

$$x_n^{k,j}, f_{(n_1,n_2)}^{k,(j_1,j_2)} \in \{0,1\} \ \ \forall k \in K, j \in V, n \in N, (j_1,j_2) \in E, (n_1,n_2) \in NE, \quad (3.20)$$

$$\sum_{(j_1,j_2) \in E} \sum_{(n_1,n_2) \in NE} f_{(n_1,n_2)}^{k,(j_1,j_2)} \cdot L(n_1,n_2) \le T_k \ \ \forall k \in K \quad (3.21)$$

$$\sum_{n\in N} {x_n}^{k,j} \sum_{i\in R} \left(\frac{1}{\frac{{d_k}^{j,i}}{{s_{n,i}}^{k,j}} - \lambda^{k,j}} \right) \leq {P_k}^{j} \quad \forall t \in T, j \in V \tag{3.22}$$

Вираз (3.14) гарантує, що для кожної ТАР/сервісу розміщується тільки одна мережева функція кожного типу. Це забезпечується тим чином, що сума змінних ${x_n}^{k,j}$ для розташування мережевої функції j повязаної з сервісом k на фізичному вузлі n по всім вузлам дорівнює одиниці.

Вираз (3.15) гарантує, що розміщення ресурсів здійснюється на фізичних вузлах, які мають адміністративну можливість для розташування відповідних мережевих функцій; оскільки гарантується, що для випадку, коли булева змінна ${x_n}^{k,j}=1$, змінна ${suit_n}^{k,j}$ буде відмінна від нуля.

Вирази (3.16), (3.17) і (3.18) являють собою обмеження на ресурси фізичних вузлів і каналів, тобто забезпечують той факт, що кількість задіяних на вузлі ресурсів не перевищує кількості наявних ресурсів. Слід зауважити, що канал між двома мережевими функціями відображається на шлях у фізичній мережі. Таким чином, його вимоги до полоси пропускання впливають не тільки на мережеві ресурси фізичних вузлів, де мережеві функції розміщуються, але й на мережеві ресурси проміжних вузлів, які лежать на шляху (вираз (3.16)).

Вираз (3.19) представляє собою обмеження щодо збереження потоку для всіх шляхів у фізичній мережі, тобто що вхідний потік на вузлі дорівнює вихідному потоку.

Вираз (3.20) гарантує, що змінні у задачі розміщення функцій мережі та відображення шляху є булевими.

Щоб обмежити затримки на каналах, обмеження на затримку, показане в виразі (3.21), також додається. А для того, щоб урахувати у моделі необхідну продуктивність віртуальної мережевої функції, у виразі (3.22) визначені обмеження на значення часу обробки заявки, що залежить від кількості виділених функціональному блоку ресурсів ${d_k}^{j,i}$ та часу обслуговування одиницею ресурсу ${s_{n,i}}^{k,j}$.

На рис. 3.9 представлено приклад системи відображення мережевих сервісів на фізичну мережу.

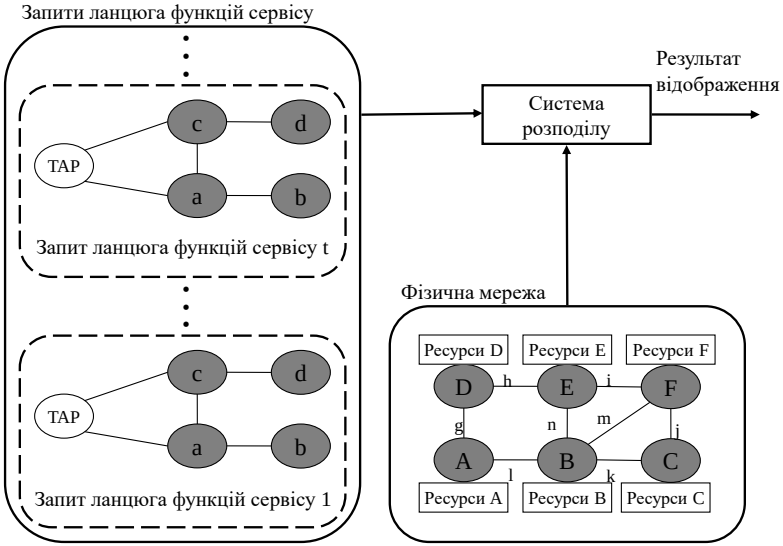


Рис. 3.9 Система виділення мережевих ресурсів – приклад топології

Передбачається вирішення задачі (3.13)-(3.22) в офлайн режимі на початковому етапі. Згідно з рішенням, кожній мережевій функції резервується певна кількість ресурсів віртуальної мережевої функції, на основі оцінки її найбільшої потреби в ресурсах; миттєві потреби різних мережевих функцій динамічно задовольняються шляхом активації необхідної конфігурації віртуальних машин під час виконання таким чином, щоб задовольнити гарантії передбачені для кожної мережевої функції.

Процес виділення ресурсів для віртуальних мережевих функцій зображено на рис. 3.10.

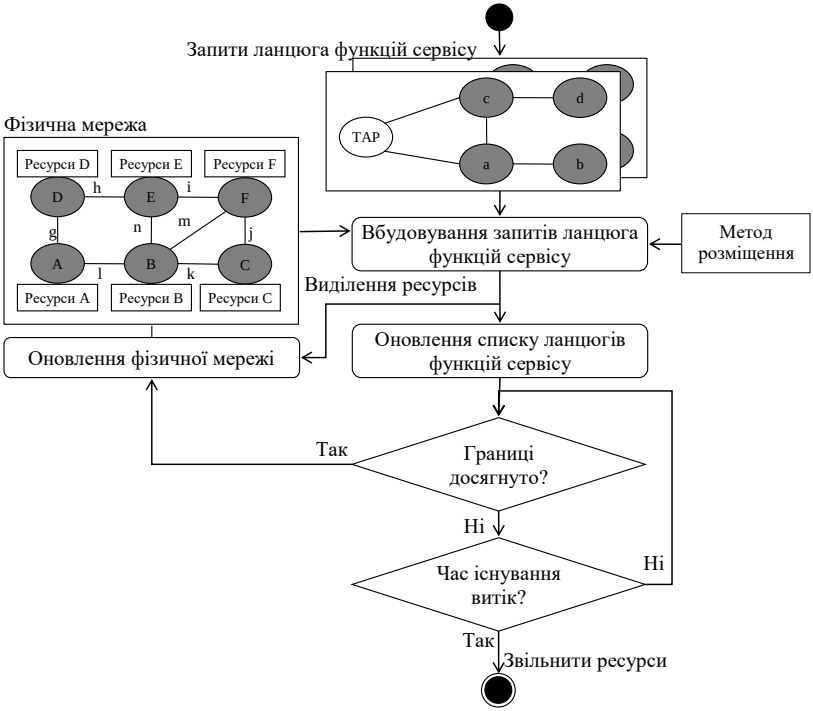


Рис. 3.10 Життєвий цикл запиту віртуальної мережі – діаграма активності

Процес виділення ресурсів починається з приходом запиту ланцюга функцій сервісу, як показано на рис. 3.10. Метод розміщення та резервування ресурсів початкового етапу (тобто (3.13)-(3.22)) використовується для вбудовування запитів ланцюга функцій мережі, він приймає на вхід поточний стан фізичної мережі (наприклад, доступні ресурси CPU, пам'яті та пропускної здатності) і самі запити.

Метод направлений на мінімізацію використання ресурсів, при цьому забезпечуючи заданий рівень якості обслуговування. Запропонований метод дозволяє телекомунікаційному оператору мінімізувати капітальні та операційні витрати використовуючи парадигму хмарних обчислень та значно покращити якість сприйняття.

3.4 Еластичне керування EPC за допомогою поділу стану

Запропонована архітектура складається з межових вузлів (Edge Node – EN) і центрального вузла (Central Node – CN). CN виступає як традиційна мережа EPC, а також надає потенційно можливості хмарних обчислень, також підтримуючи глобальне

уявлення про систему. Мережеві функції можуть розгортатися як в CN і так і в вузлах EN, користуючись усіма перевагами, які надходять від хмарних механізмів [103].

Основна ідея модифікованого методу полягає в ієрархічно розподіленій архітектурі (рис. 3.11), що дозволяє мінімізувати навантаження на мережу, зменшує навантаження на пристрої (процесор, пам'ять і енергію) і прискорює процес функціонування. Для цього вводиться проксі менеджер синхронізації (Proxy Synchronization Manager – PSM). PSM виступає в якості проксі сутності для переговорів з іншими сутностями мережі, зокрема, обслуговуючим менеджером синхронізації (Serving Synchronization Manager - SSM) і домашнім менеджером синхронізації (Home Synchronization Manager - HSM). SSM масштабується відповідно до очікуваних сервісів мережі і містить інформацію стану стосовно сусідніх EN. SSM передає інформацію управління до PSM. HSM є центральним вузлом, який зберігає інформацію щодо стану всієї мережі. HSM і SSM є ключовими елементами для управління великою кількістю вузлів і вони мають вирішальний вплив на продуктивність мережі.

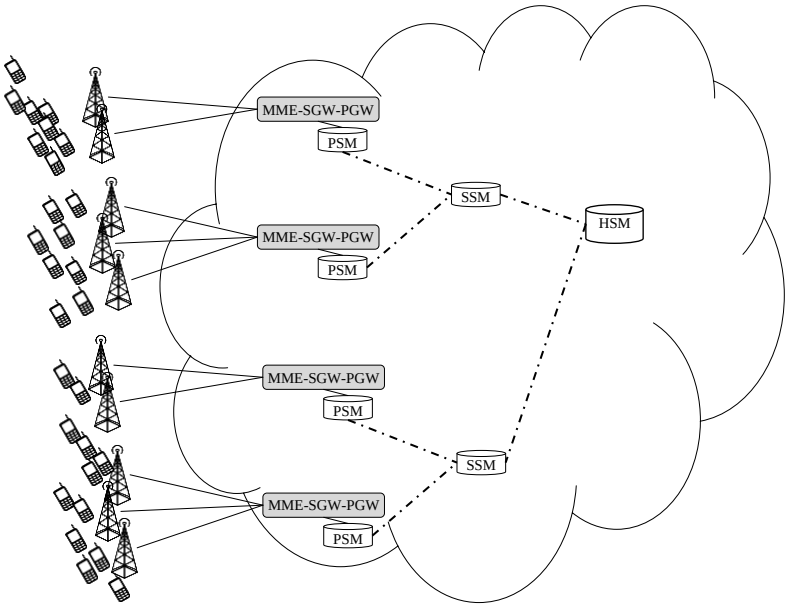


Рис. 3.11 Мережна архітектура з системою керування станом

Розглянемо тепер необхідність покращеної координації для забезпечення передачі стану між EN. Не представляється можливим використовувати існуючі процедури EPC і повинні бути введені деякі додаткові механізми. Серед кількох можливих альтернативних рішень, ми вибрали для поділу станом проактивну передачу стану і

синхронізацію зміни стану у сусідніх EN випереджувальним чином, всякий раз, коли нова стан з'являється в системі або оновлюється (рис. 3.12).

Переваги такого підходу різноманітні. По-перше, кожен компонент завжди володіє оновленою інформацією про всіх користувачів і стан синхронізований і узгоджений по всій системі, що дозволяє компонентам бути прозоро доданими або видаленими. Крім того, логіка кожного елементу не знає, що інші компоненти одного і того ж типу знаходяться в системі; що призводить до практично площини управління, що не зберігає стан, в той час як кожен з компонентів відслідковує стан і може бачити повну інформацію про стан в системі. Зосереджуючи увагу на відмовостійкості, можливість легко реплікувати стан в різних надлишкових компонентах дозволяє уникнути ситуації, що стан втрачається в разі відмови компонента і видаляє єдині точки відмови. Нарешті, процедура синхронізації стану відповідає стандарту EPC, вона по своєму дизайну є прозорою до процедур EPC і не вимагає змін у стандартних протоколах зв'язку між компонентами EPC, таким чином, забезпечує швидке впровадження і інтеграцію з існуючими реалізаціями.

Відзначимо, що введення політики балансування і реплікації стану в EN дає можливість будь-якому EN обробляти запити, що надходять від того ж абонента: eNodeB може виконувати операції балансування навантаження з дрібним рівнем та рівнем запитів.

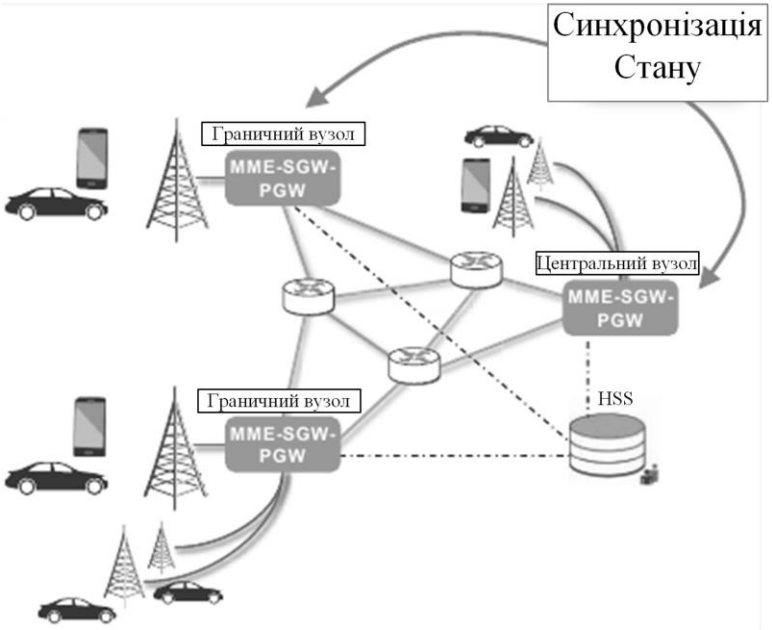


Рис. 3.12 Синхронізація стану EPC

Для того, щоб мати можливість транспортувати інформацію про стан контрольованим чином через мережу, існує потреба в розробці специфічного інтерфейсу і афферентного комунікаційного протоколу. Такий інтерфейс повинен бути визначений в контексті 3GPP, а також в більш загальному контексті, дозволяючи будь-який тип передачі стану.

Далі опишемо Межовий Протокол Синхронізації (Edge Synchronization Protocol – ESP) у вигляді набору різних процедур для координації вузлів EN. Перша процедура називається Процедура Встановлення Інтерфейсу Синхронізації і її основною метою є встановлення початкового загального стану абонента між двома або більше сутностями EN. Друга називається Процедура Повторної Синхронізації, забезпечує належну повторну синхронізацію двох різних EN, коли зв'язок між ними навмисно або ненавмисно втрачається через різні збережені стани, коли підключення повертається. Протокол може бути використаний між декількома EN в разі стандартних реалізацій EPC 3GPP, а також для будь-якої інформації про стан абонента, яка має формат, аналогічний телекомунікаційному.

3.4.1 Процедура Встановлення Інтерфейсу Синхронізації

Ця операція відбувається, коли новий EN під'єднується до пулу [103]. Припускаючи, що EN1 є вузлом, що приєднується, пропонована процедура синхронізації починається з того, що EN1 звертається з проханням приєднатися до пулу синхронізації стану абонента, де EN2 вже приєднаний. Як показано на рис. 3.13, перша фаза процедури починається з прив'язки між двома об'єктами. EN1 підписується на зміни у інформації стану, з точки зору нової або оновленої, і, навпаки, EN2 робить підписку на модифікації, які відбуваються в EN1 (кроки 1 і 2). Після чого, вони можуть обмінюватися один з одним станами абонентів, які вони в даний час зберігають. Цей крок (крок 3) також виконується щоразу, коли стан оновлюється або приєднується новий абонент, таким чином, після стандартних процедур Attach, Detach та Handover. Крім того, щоб забезпечити узгодженість станів в системі, кожен стан негайно поділяється після завершення процедури.

У разі відключення між контролерами, повинна бути виконана процедура синхронізації [103]. Як пояснено нижче, спроектоване рішення засноване на часових мітках, і, таким чином, вимагає синхронізації годинників прилеглих вузлів EN. Використовується Network Time Protocol (NTP), стандартний протокол для синхронізації годинників на мережах з пакетною комутацією та змінної затримкою мереж [104].

Розглядаючи тепер процедуру повторної синхронізації, вона полягає в перевірці обміну часовими мітками вузлів EN кожен раз, коли стан пересилається, в повідомленні «SubscriberStateTransmission», так що під час фази повторного з'єднання кожен EN може знати, який з них володіє новим станом. Приклад описаний на рис. 3.14, в якому під час відключення двох EN, UE модифікує абонентський стан за допомогою процедури, що виконується з EN2 (крок 3). Коли вони стають знову підключені, EN1 посилає стан абонента разом з відповідною міткою часу (TS1) (крок 4). Оскільки TS1 старіше TS2, EN2 відправляє назад новий стан і EN1 оновлює свій власний (кроки 5 і 6).

Рис. 3.13 Процедура встановлення інтерфейсу сигналізації [103]

3.4.2 Процедура Повторної Синхронізації

Підкреслимо, що введені розширення є прозорими для базової мережі 3GPP; іншими словами, описані розширення інтегруються в EPCaaS, підтримуючи його сумісним зі стандартом, в той же час використовуючи прозорі механізми.

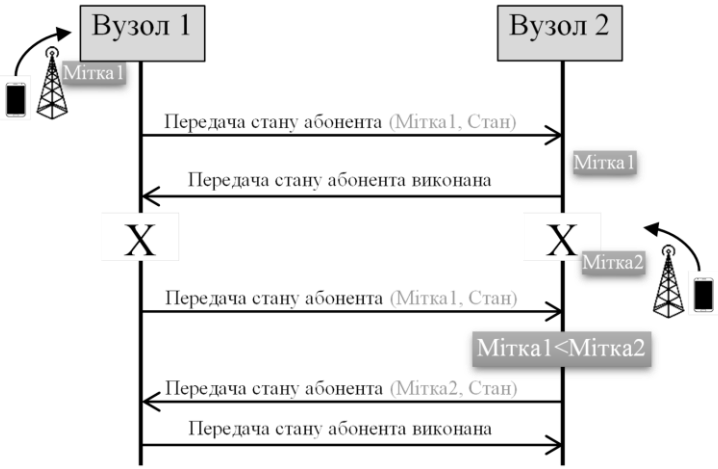


Рис. 3.14 Процедура повторної синхронізації [103]

3.5 Оркестрування віртуалізованої мережі EPC

Загальна архітектура системи керування гетерогенною мережею представлена на рис. 3.15.

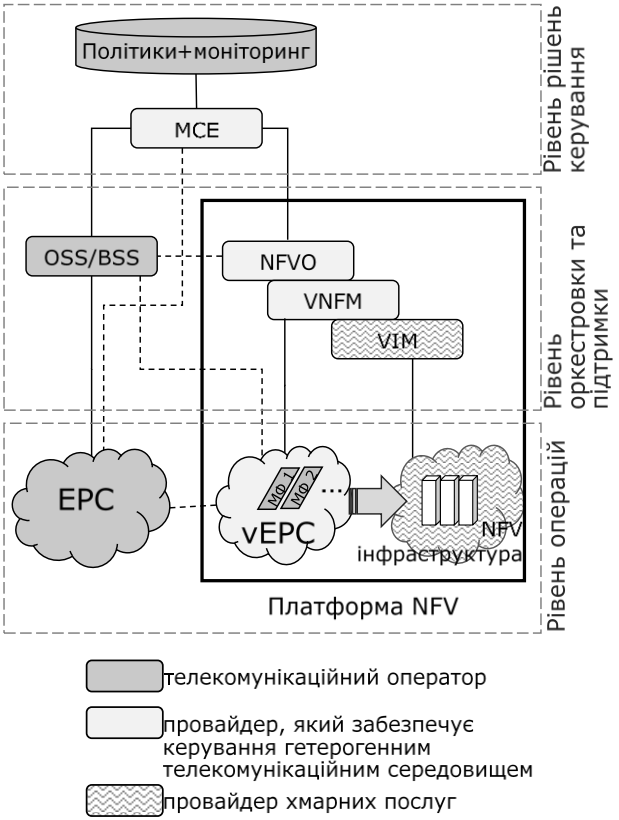


Рис. 3.15 Архітектура системи керування гетерогенною мережею

Процедура створення віртуалізованої базової мережі (vEPC) за запитом починається з того, що подія моніторингу посилається від традиційної мережі LTE EPC (LEPC) у напрямку до OSS/BSS (крок 1), як показано на рис. 3.16 [105].

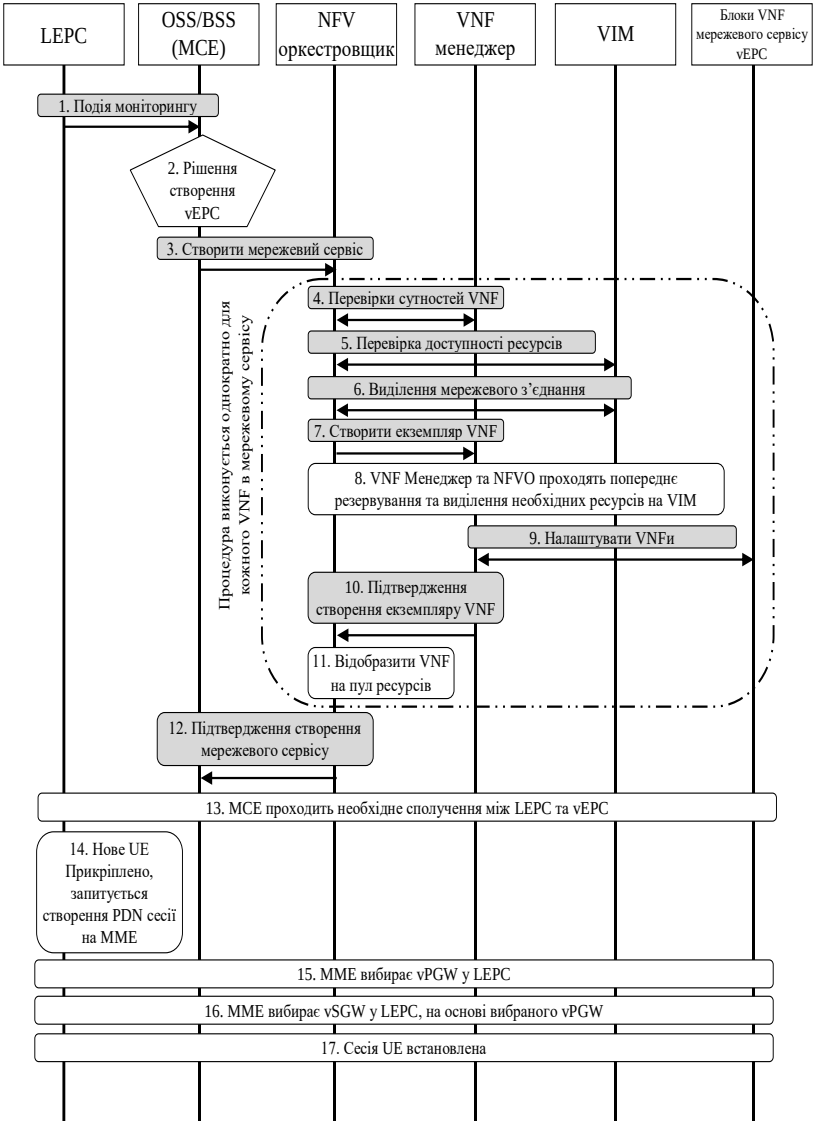


Рис. 3.16 Потік повідомлень створення сутностей VNF для vEPC

Цей приклад ілюструє специфічним випадок режиму моніторингу OSS/BSS, де сутності здатні збирати метрики мобільних мереж від різних елементів EPC, оцінюючи операційний стан мережі і повідомляючи його в блок керування Management Component Entity (MCE). За допомогою цієї інформації, MCE (інтегрований в OSS/BSS для спрощення рисунку) може прийняти рішення про створення сутностей віртуалізованного мережевого сервісу EPC, дозволяючи оператору вирішувати вимоги зростання сервісу (крок 2). Таким чином, команда створення сутностей відправляється до NFVO (крок 3). Це ініціює процедуру, яка повторюється для кожного VNF, що складає NS (що формується через граф передачі VNF), що починається з перевірки NFVO існуючих екзмеплярів VNF в VNFM (крок 4). Це дозволяє йому визначити, які нові VNF повинні бути створені і для кожного VNF NFVO запитує VIM наявні ресурси (крок 5), а також запитує виділення мережевих з'єднань між VNF, які є частиною мережевого сервісу, над віртуалізованим мережевим ресурсом (крок 6). Після завершення попередніх етапів, умови збираються NFVO для запиту фактичного створення екземплярів сутностей окремих VNF до VNFM (крок 7), що охоплює попереднє резервування і виділення необхідних ресурсів на VIM (крок 8). VNFM переходить до конфігурації конкретних VNF, відповідно до їх специфікації дескрипторів або конкретних параметрів розгортання (крок 9). У разі якщо VNF вимагає сполучення з блоком Управління Елементом (Element Management), цей крок конфігурації дозволяє VNFM повідомити EM, який потім може застосувати параметри специфічні для прикладної програми до VNF після їх створення. VNFM потім відповідає назад NFVO про їх успішну фіналізацію створення сутностей (крок 10), і далі відображення на пул ресурсів (крок 11). Потім процедура підтверджує наявність мережевого сервісу зі сторони MCE (крок 12), який потім може приступити до необхідного сполучення між фізичними елементами LEPC, і їх віртуалізованими відповідностями, які знаходяться в vEPC (крок 13). Тобто, HSS, PCRF та MME інформуються про створену vEPC. Таким чином, відповідні інформаційні контексти стосовно наявних vPGW та vSGW оновлюються в цих сутностях.

Коли UE приєднується до LEPC (крок 14), запит на приєднання UE обробляється в MME, який виконує вибір PGW та SGW для приєднаного UE, виконуючи стандартну процедуру 3GPP. Передбачається, що вагові коефіцієнти IP-адрес vPGW/vSGW, повернені з DNS-сервера, перевищують значення інших IP-адрес PGW/SGW. Для цього, як тільки мережа vEPC буде створена і готова до сервісу vEPC, нові DNS записи створених vPGW/vSGW повинні бути додані до внутрішнього DNS OSS/BSS.

Приклад робочого процесу оркестрації сервісу представлено на рис. 3.17.

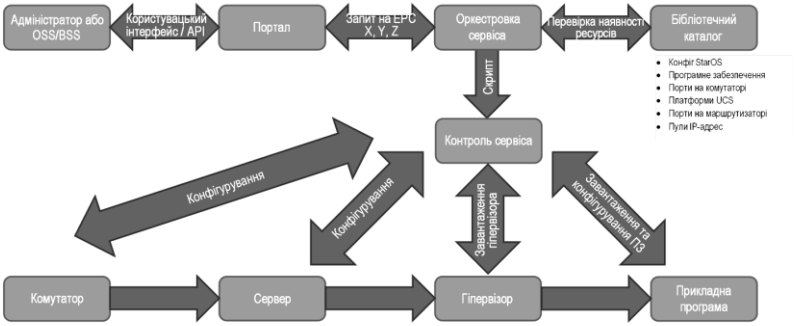


Рис. 3.17 Приклад робочого процесу оркестрації сервісу [106]

3.1 Оцінка методу визначення місця розміщення та необхідної ємності віртуальних ресурсів у разі виникнення перевантаження фізичної мережі

Вирішення задачі (3.13)-(3.22) здійснювалось з використанням генетичного алгоритму у системі MATLAB. Фрагмент коду програми представлено на рис. 3.18.

```
>>ObjectiveFunction=@fitness;
>>nvars=108;% Number of variables
>>LB=zeros(1,108);% Lower bound
>>UB=ones(1,90);% Upper bound
>>IntCon=ones(1,90);% Integer variables
>>ConstraintFunction=@constraint;
>>[x,fval]=ga(ObjectiveFunction,nvars,[],[],[],[],LB,
UB,ConstraintFunction,IntCon,options)
```

Рис. 3.18 Фрагмент коду MATLAB

Приклад моделювання системи з десятьма вузлами, трьома функціональними блоками та двома типами ресурсів показав зменшення умовних витрат в середньому до 15% при використанні методу визначення оптимального місця розміщення ресурсів у порівнянні зі стратегією коли EPC як сервіс надається єдиним наперед визначеним хмарним вузлом (рис. 3.19).

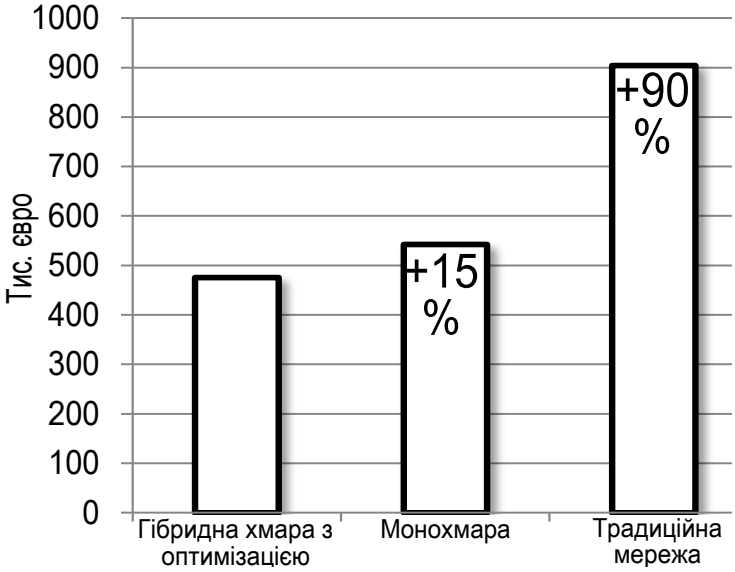


Рис. 3.19 Загальна вартість розміщення

Висновки

1. Запропоновано метод керування розміщенням та визначенням необхідного обсягу ресурсів мережевих функцій для оптимізації кількості ресурсів, виділених мережевій функції у мережі телекомунікаційного оператора.
2. Запропонований метод дозволяє пов’язати продуктивність мережевої віртуалізованої функції з виділеною їй кількістю ресурсів, а також враховує гетерогенність телекомунікаційного середовища.
3. Розглянуто архітектурне рішення еластичного керування EPC за допомогою поділу стану, а також процес взаємодії між мережними елементами від час розгортання віртуалізованих мережних вузлів.
4. Метод може застосовуватись при керуванні розгортанням мережевих функцій у гетерогенному апаратному середовищі для мінімізації витрат оператора зв'язку та покращення якості обслуговування абонентів.

Контрольні запитання

1) Що таке абонентське обладнання (UE)?
2) Опишіть основні процедури сигналізації.
3) Що таке потік заявок?
4) В чому полягає задача вибору вузлів агрегації трафік ?

5) Які цілі оптимізації та обмеження в задачі вибору вузлів агрегації трафіку?
6) В чому полягає задача відображення віртуальних вузлів на фізичні вузли?
7) Поясніть еластичне керування EPC за допомогою поділу стану.
8) Як проходить процедура встановлення інтерфейсу синхронізації?
9) Як проходить процедура повторної синхронізації?
10) Опишіть архітектуру системи керування гетерогенною мережею.
11) Поясніть потік повідомлень створення сутностей VNF для vEPC.
12) Опишіть архітектуру системи керування гетерогенною мережею.
13) Поясніть потік повідомлень створення сутностей VNF для vEPC.

РОЗДІЛ 4
СИСТЕМА КЕРУВАННЯ РЕСУРСАМИ ВІРТУАЛІЗОВАНИХ МЕРЕЖЕВИХ ФУНКЦІЙ

4.1 Постановка задачі

Метою системи керування ресурсами мережевих функцій є виділення достатньої їх для задоволення вимог до якості обслуговування навантаження. В основі будь-якого алгоритму надання ресурсів лежать два питання: скільки надавати і коли [84].

Скільки надавати: Для вирішення питання про те, скільки ресурсів виділити для кожної мережевої функції, використовуємо модель серверу з розділенням процесору. Модель приймає в якості вхідних даних інтенсивність надходження вхідних запитів і вимоги обслуговування окремого запиту, і обчислює кількість ресурсів, необхідних мережевій функції, щоб впоратися з вимогами.

Коли надавати: Рішення про те, коли надавати ресурси, залежить від динаміки навантажень. Телекомунікаційні навантаження зазнають довгострокових змін, таких як вплив години дня або сезонні ефекти, а також короткострокових коливань таких як несподівані натовпи. У той час як довгострокові коливання можуть бути передбачені заздалегідь, спостерігаючи за змінами в минулому, короткострокові коливання менш передбачувані, а в деяких випадках, не передбачувані. Запропонована методика використовує два різних методи для роботи в умовах змін, які спостерігаються в різних часових масштабах. Використовується прогностичне управління ресурсами для оцінки навантаження і відповідного керування, а також реактивне керування ресурсами для реагування на непередбачені зростання у кількості заявок, що надходять.

Розглянемо мережу, в якій функціонує кілька мережевих функцій [73]. Передбачається, що кожна така мережева функція вказує бажану вимогу до якості обслуговування (QoS); при цьому в даному випадку передбачаємо, що вимоги до QoS визначені в термінах цільового часу відповіді. Метою системи є забезпечення того, що середній час відповіді (або деякий процентиль часу відповіді), який спостерігається запитами мережевої функції не перевищує бажаний цільовий час відповіді. Загалом, кожен вхідний запит обслуговується декількома апаратними та програмними ресурсами на сервері, такими як CPU, NIC, диск і т.д. Припускаємо, що заданий цільовий час відповіді розділяється на кілька значень часу відповіді для конкретних ресурсів по одному для кожного такого ресурсу. Таким чином, якщо кожен запит на кожному ресурсі не витрачає часу більше, ніж призначене цільове значення, то загальний цільовий час відповіді для сервера буде задоволений. Задача розділення зазначеного значення часу відповіді сервера на значення часу відповіді для конкретного ресурсу виходить за рамки даного дослідження; у роботі передбачається, що такі конкретні для ресурсу значення часу відповіді задані.

Системи обслуговування, що представляють собою ресурси (процесор та пам’ять), передбачаються послідовними, тобто обслуговування протікає конвеєром. Додатково, приймається дисципліна обслуговування з розділенням процесора (processor sharing – PS) у кожній черзі, оскільки PS апроксимує політику диспетчеризації, що

використовується більшістю операційних систем, наприклад, зважену справедливу організацію черги та розподіл машинного часу у Linux. Для простоти подальшого викладу припускаємо, що в системі наявний лише один тип ресурсу.

Формально, позначими цільовий час відповіді мережевої функції i як P_i і T_i – спостережуваний середній час відповіді, тоді мережевій функції потрібно виділити таку кількість ресурсів, щоб $T_i \leqslant P_i$.

Використаємо таку постановку задачі щоб одержати механізм динамічного виділення ресурсів, який описано далі.

4.2 Система динамічного керування ресурсами

Щоб виконати динамічне виділення ресурсів, засноване на вищевказаному формулюванні, на кожному сервері необхідно буде використовувати три компоненти: (1) модуль моніторингу, який вимірює навантаження і показники продуктивності кожної мережевої функції (такі як інтенсивність надходження запитів, середній час відповіді тощо), (2) модуль прогнозування, який використовує вимірювання з модуля моніторингу для оцінки характеристик навантаження в найближчому майбутньому, і (3) модуль розподілення ресурсів, який використовує ці оцінки навантаження для визначення кількості ресурсів, яку необхідно виділити мережевим функціям. На рис. 4.1 показані ці три компоненти [73].

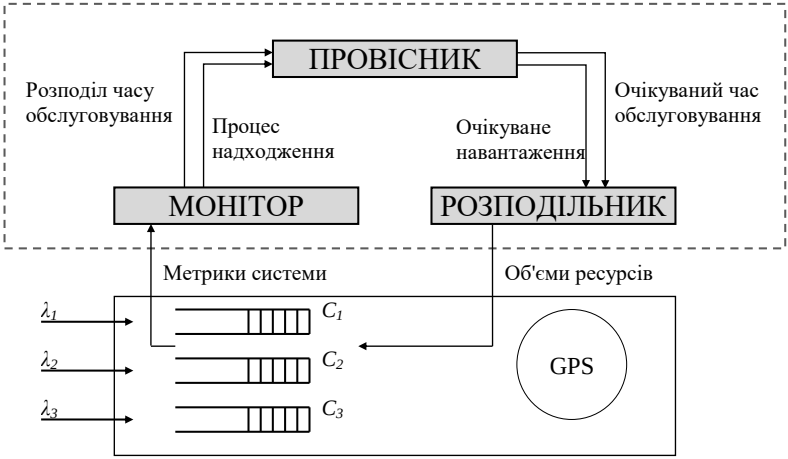


Рис. 4.1 Динамічне виділення ресурсів

Як вже відмічалося, навантаження часто мають передбачувані шаблони. Проте, можуть бути значні відхилення від цих моделей через пульсуючий характер навантажень. Надлишкові заявки або раптові сплески запитів можуть привести до погіршення обслуговування. Крім того, існують витрати та ризики пов’язані зі змінами у системі, і,

таким чином, прагнуть уникнути частих змін конфігурації ресурсів. Грунтуючись на цих спостереженнях, пропонується гібридне рішення системи керування виділенням ресурсів, яке поєднує в собі прогностичну складову з реактивною.

На рис. 4.2 показана концептуальна архітектура пропонованого рішення:

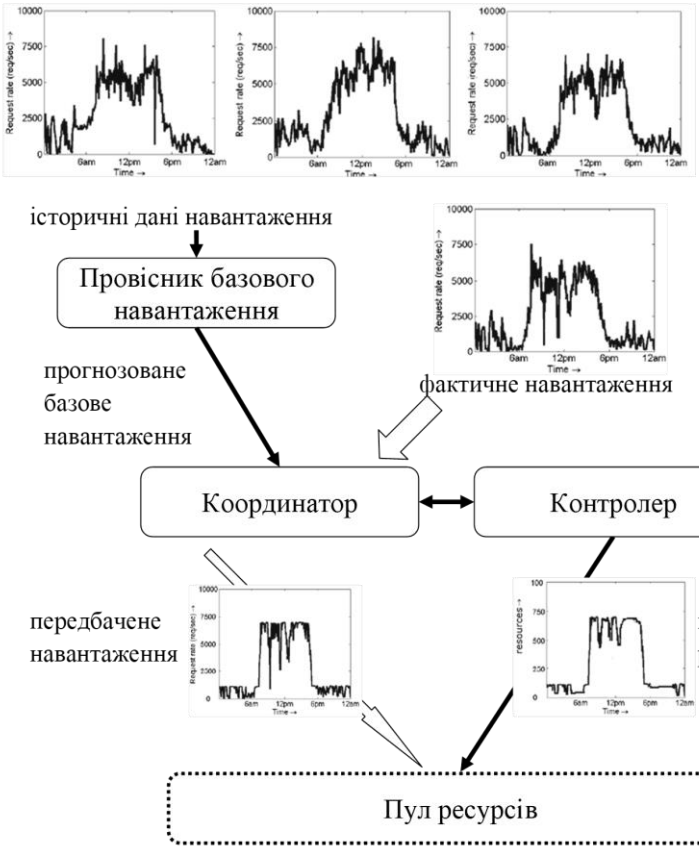


Рис. 4.2 Система гібридного керування

1. Провісник базового навантаження аналізує історичні дані навантаження і визначає закономірності, які формують базове навантаження.
2. Координатор направляє запити навантаження на сервери, а також обмінюється даними з контролером для надання інформації про вхідне навантаження.
3. Контролер оцінює і виділяє відповідну кількість ресурсів необхідну для обробки навантаження. Він генерує розподіл ресурсів (кількість ресурсів на рис. 4.2).

4.3 Моніторинг та керування

Модуль онлайн моніторингу відповідає за вимірювання різних метрик системи і мережевих функцій. Ці показники використовуються для оцінки параметрів моделі системи і характеристик навантаження.

Моніторинг на мережі реалізується встановленням головної прикладної програми моніторингу та відповідних підконтрольних їй агентів на кожен мережевий блок [85].

Важливий аспект платформи полягає в здатності моніторингу її різних складових елементів [105]. Цей механізм дозволяє платформі адаптувати і масштабувати мережеві функції у відповідності з різними метриками, де сутності прийняття рішень порівнюють отримані тригери з вимогами описів сервісів окремих елементів, які повідомляються ним. Платформа забезпечує можливість відділення аспектів моніторингу функціонування мережі мобільного зв'язку від тих, які відносяться до інфраструктури NFV, підтримуючи існуючі механізми керування та оркестровки в зв'язку з цим. Конкретно, інтерпретація подій моніторингу OSS/BSS і MCE з фізичної LTE EPC може бути використана для запуску створення віртуальної EPC (vEPC) в інфраструктурі NFV, або навіть просто створення одного елемента vEPC або компонента. Таким чином, доступні механізми моніторингу в платформі надаються на різних рівнях:

- VNF: коли VNF вбудовує функції моніторингу і передає зібрану інформацію в Element Management (EM) або VNF Manager (VNFM). Система дозволяє функціям VNF безпосередньо направляти дані сутностям OSS/BSS і MCE з подіями моніторингу, що дозволяє їм приймати загальномережеві управлінські рішення, з урахуванням також віртуалізованих функцій;
- EM: EM може реалізувати рішення про дії, на основі даних, зібраних з VNF, і направити їх на VNFM, або до OSS/BSS і MCE для дій на всю мережу;
- OSS/BSS і MCE: коли функція моніторингу не реалізована в EM, або перетинає кілька EM. Це може бути як точкою детектування, так і рішення. OSS/BSS також отримує події моніторингу LTE EPC, що дозволяє запускати зміни в віртуалізованій мережі мобільного зв'язку або її елементів;
- Оркестровщик NFV (NFVO): коли робота віртуалізованного сервісу зазнає впливу на рівні мережевого сервісу (тобто складається з декількох взаємопов'язаних VNF з загальним кінцем), то NFVO може бути покращено можливістю спільно розглядати різні звіти про моніторинг окремих VNF, і діяти, коли функціональні потреби мережевого сервісу в цілому піддаються впливу, запускаючи VNFM діяти над необхідними VNF;
- VNFM: він приймає рішення масштабування, основуючись на подіях VNF, відповідно до параметрів масштабування і моніторингу в їх дескрипторах. Цей об'єкт може використовувати цю інформацію, а також діяти у відношенні необхідних дій інфраструктури NFV, отримуючи події від менеджера віртуалізованої інфраструктури (Virtualized Infrastructure Manager – VIM).

Таким чином, результат моніторингу в розглянутій платформі спрямований в бік загальної поведінки масштабування віртуальних ресурсів відповідно до продуктивності мережевого сервісу/VNF, трьома шляхами: 1) автоматичне масштабування, де VNFM моніторить події і запускає операцію масштабування при виконанні певних умов, 2) масштабування на вимогу, коли VNF, або його EM, викликає операцію масштабування за

допомогою явного запиту на VNFM, або 3) запит керування, який ініціюється відправником (OSS/BSS або MCE) у напрямку до VNFM через NFVO.

Для того, щоб здійснювати моніторинг роботи площини користувача і площини управління, S1-U і S5 є точками спостереження для моніторингу продуктивності трафіку даних, в той час як S1-MME і S10 є точками спостереження для моніторингу продуктивності навантаження сигналізації через необхідні операції і активність абонентів для користувальницького обладнання. Інформація моніторингу з означених опорних точок спостереження збирається на OSS/BSS. Інформація від сутностей компонентів vEPC збирається на кожному EM і доставляється зрештою OSS/BSS. Система дозволяє мережевим функціям безпосередньо направляти дані сутностям OSS/BSS і MCE з подіями моніторингу, що дозволяє їм приймати загальномережеві управлінські рішення, з урахуванням також віртуалізованих функцій.

Диференціація інтенсивності моніторингу окремих сегментів мережі дає можливість усунути надлишкові процеси моніторингу та обробки даних, забезпечуючи необхідну інтенсивність лише на тих вузлах, де це необхідно. Використовуючи такий підхід можливо підвищити гнучкість управління мережевими функціями, і, як наслідок, ефективніше використовувати ресурси пристроїв рівня управління. Крім того, володіючи актуальною службовою інформацією, можна здійснювати оптимальне керування навантаженням у мережі. Це дозволить значно зменшити витрати на електроенергію та збільшити час «життя» елементів мережі.

4.4 Визначення інтервалу часу сталої конфігурації ресурсів

Оскільки мережевий трафік не є статичним це може привести до неефективного використання загальних ресурсів протягом дня, з одного боку. З іншого, необхідно забезпечувати задану якість обслуговування. Тому пропонується стратегія коли конфігурація віртуальних машин створюється завчасно з розгляду найгіршого випадку, а активізується коли виникає необхідність.

Пропонується застосовувати метод динамічного визначення розміру інтервалу часу сталої конфігурації ресурсів, тобто визначення періодів часу протягом яких конфігурація ресурсів буде незмінною – часу між двома послідовними запусками алгоритму реконфігурації. Суть цього механізму полягає у динамічній зміні інтенсивності здійснення перерозподілу ресурсів залежно від різниці між мінімальним та максимальним значенням навантаження мережевої функції на інтервалі на основі зібраних за довготривалий період часу даних стану елемента мережі. Якщо це значення збільшується, то скорочується інтервал сталої конфігурації мережевих функцій.

Формула (4.1) описує принцип зміни інтервалу часу сталої конфігурації ресурсів:

$$Int(t) = \\ = max\left(\begin{matrix} Int_{base} \cdot \left(1 - K \cdot \frac{\max_{\tau\epsilon(t-I(t-1);t)} \lambda_{basepred}(\tau) - \min_{\tau\epsilon(t-I(t-1);t)} \lambda_{basepred}(\tau)}{max\lambda_{basepred}}\right); \\ Int_{minbase} \end{matrix}\right), \quad (4.1)$$

де *Int* – інтервал часу сталої конфігурації ресурсів, Int_{base} – базове значення інтервалу, $Int_{minbase}$ – мінімальна допустима величина базового інтервалу, *K* – коефіцієнт зміни тривалості сталої конфігурації, $\lambda_{basepred}(t)$ – середньостатистична передбачена інтенсивність надходження заявок під час інтервалу *t*.

Визначаємо і дискретизуємо шаблони у прогнозованому навантаженні. Мета полягає в тому, щоб представити денний шаблон в навантаженні дискретизуючи його запити у послідовні, непересічні часові інтервали з єдиним репрезентативним значенням вимоги в кожному інтервалі. Запропоновано алгоритм для знаходження невеликої кількості часових інтервалів, таких, що відхилення від фактичної вимоги зведено до мінімуму. Підтримувати невелику кількість інтервалів важливо, так як більше число інтервалів означає більш часті зміни виділення ресурсів і, таким чином, більш високі ризик і витрати. Визначимо формально дискретизацію [57].

Дискретизація навантаження: Маючи часовий ряд X на області $[v, \tau]$, часовий ряд Y на тій же самій області є характеризацією/дискретизацією навантаження X якщо $[v, \tau]$ може бути розділена на m послідовних непересічних часових інтервалів, $\{[v, \tau_1],[\tau_1, \tau_2],\ldots,[\tau_{m-1}, \tau]\}$,так що $X(j)=r_i$, для всіх j у $i^{му}$ інтервалі, $[\tau_{i-1}, \tau_i]$.

Зверніть увагу, за визначенням, будь-який часовий ряд, X, є дискретизацією сам по собі. Встановлюємо v=0 і τ=період навантаження. У подальшій дискусії припускаємо період у 24 години, на основі аналізу періодичності.

Ідея дискретизації подвійна. По-перше, потрібно точно відображати вимоги. Для досягнення цієї мети, репрезентативні значення, r_i, для кожного інтервалу, $[\tau_{i-1}, \tau_i]$, мають бути якомога ближчими до фактичних значень часового ряду в інтервалі $[\tau_{i-1}, \tau_i]$, По-друге, виділення ІТ-ресурсів не відбувається безоплатно [14]. З цієї причини, потрібно уникати занадто великої кількості інтервалів і, отже, занадто великої кількості змін в системі, так як це не практично і може призвести до багатьох проблем (наприклад, втрати продуктивності, зносу серверів, нестабільності системи і т.д.) , Таким чином, слід мінімізувати помилку, внесену дискретизацією, і кількість інтервалів в дискретизації. Пропонується рішення для дискретизації часових рядів таким чином, щоб зменшити обидві величини.

$$w_1 \cdot f_1(m) + w_2 \cdot f_2(m) \rightarrow min \tag{4.2}$$

Вираз (4.2) є цільовою функцією яку хочемо мінімізувати, де w_1, w_2 –коефіцієнти нормалізації, $f_1(m)$ є функцією відхилення репрезентації від кривої навантаження при m інтервалах (наприклад, лінійна), $f_2(m)$ є функцією вартості кількості змін (інтервалів m).

$$f_1(m) = \sum_{i=1}^{m} \left[\sum_{\tau=\tau_{i-1}}^{\tau_i} u\left(r_i - \lambda_{basepred}(\tau)\right)\right] \tag{4.3}$$

$$r_i = \max_{\tau\epsilon(\tau_{i-1},\tau_i)} \lambda_{basepred}(\tau) \tag{4.4}$$

Метою виразу (4.2) є одночасна мінімізація помилки репрезентації навантаження і кількості змін.

У деяких випадках, можна було б віддати перевагу, мінімізації квадрату кількості змін (або будь-якій іншій функції кількості змін). Вибір кращої функції вартості повинен здійснюватися мережевими адміністраторами з урахуванням конкретних потреб мережі під управлінням [107]. Для даної роботи встановлюємо $f_2(m)=l*m$, де l є константою нормалізації. Цільова функція виражає мету, мінімізуючи нормалізовану кількість змін і помилку репрезентації. Функція вартості репрезентативної помилки використовується для кількісної оцінки похибки подання навантаження. Зверніть увагу, що в найзагальнішому випадку, як кількість змін, так і репрезентативна помилка можуть бути сформульовані як функції корисності.

Функція вартості кількості змін, наприклад, може відображати вартість електроенергії, а також грошову вартість пов'язану зі зносом на перезавантаження, яка

базується приблизно на вартості заміни диска і номінальному часі напрацювання на відмову [108].

Оптимальну величину базового інтервалу визначаємо задаючи різні значення кількості інтервалів та обчислюючи значення виразу (4.2) і обираючи найкраще, тобто мінімальне, при цьому:

$$Int_{base} = \frac{\tau_m}{m} \tag{4.5}$$

Приклади репрезентації значень часового ряду представлені на рис. 4.3, де помилка репрезентації для випадку інтервалів у 10 хвилин складає 7%, а для випадку 60 хвилин – вже 19%.

Рис. 4.3 Вплив величини інтервалу сталої конфігурації ресурсів на точність представлення очікуваного навантаження

Псевдокод алгоритму (рис. 4.4) буде мати вигляд, представлений на рис. 4.5.

Рис. 4.4 Визначення базового інтервалу

```
interval←maxinterval
i_interval←0
startpoint←begining
while interval>0
  startpoint←0
  endpoint←startpoint+interval
  while startpoint<ending
        maxvalue←max[Load(startpoint)… Load(endpoint)]
        for j∈startpoint... endpoint
              BaseLoad_interval(j)←maxvalue
              DiffLoad_interval(j)←BaseLoad_interval(j)-Load(j)
        startpoint←endpoint
        endpoint←startpoint+interval
        i_interval←i_interval+1
  interval←interval-1
for j∈1...maxinterval
  Minfunc_j←∑_k DiffLoad_j(k)+l* i_j
MinInt←argmin(Minfunc)
MinCost←min(Minfunc)
```

Рис. 4.5 Псевдокод алгоритму визначення оптимального базового інтервалу

З іншого боку, додатково можем ввести умову, що різниця між сусідніми репрезентативними значеннями не повинна бути меншою певного заданого порогу. Тобто для всіх інтервалів проводиться аналіз r_i, якщо$|r_i-r_{i+1}|<\varepsilon$, де ε – деякий поріг, тоді значення i та $(i+1)$ обєднуються, а нове репрезентативне значення обирається як $max(r_i;r_{i+1})$.

4.5 Розподілення ресурсів мережевих функцій

Опишемо модель, яка використовується для визначення потреби в ресурсах мережевої функції на основі очікуваного нею навантаження і цільового часу відповіді.

Оскільки ресурс моделюється як GPS сервер, інтенсивність обслуговування запиту мережевої функції дорівнює d/s, де d – це кількість ресурсів мережевої функції і s – це середній час обслуговування запиту однією одиницею ресурсу. Отже, інтенсивність обслуговування запиту

$$\mu = d/s, \tag{4.6}$$

Засновуючись на теорії масового обслуговування, розподіл часу обслуговування заявок такої системи є експоненційними з середньою величиною $\frac{1}{\mu-\lambda_{basepred}}$. Мережевій

функції потрібно виділити кількість ресурсів, так що $T \leqslant P$. Цей вираз можна переписати як

$$T = \frac{1}{\mu - \lambda_{basepred}} \rightarrow d_i \geq s\left(\frac{1}{T} + \lambda_{basepred}\right) \quad (4.7)$$

4.6 Метод прогнозування навантаження

Визначення конфігурації ресурсів описане в попередньому пункті у значній мірі залежить від точної оцінки навантаження, що може надійти. Пропонується здійснювати постійний моніторинг значень інтенсивності навантаження та використовувати прогностичний метод, якщо ці значення не перевищують порогові, в іншому випадку – оцінювати поточні тренди у навантаженні та здійснювати масштабування ресурсів на основі нового прогнозу.

У цьому пункті представляємо метод прогнозування у разі несподіваних відхилень від початкового прогнозу, що використовує минулі спостереження для оцінки майбутнього навантаження мережевої функцій.

Найважливішим параметром, який характеризує вхідне навантаження, є інтенсивність надходження запитів, і точне її визначення дозволяє оцінювати середній час перебування заявки у системі.

Провісник навантаження використовує минулі спостереження за навантаженням, щоб передбачити пікову вимогу, яка буде зазнаватися протягом часу *I*. Для простоти викладу припустимо, що *I* = 1 година. Щоб зробити це модуль прогнозування зберігає історію інтенсивності надходження заявок, яка мала місце протягом кожної години дня, протягом останніх декількох днів. Потім для кожної години генерується гістограма, використовуючи спостереження для цієї години за останні декілька днів (див. рис. 4.6). Кожна гістограма дає розподіл ймовірностей інтенсивності надходження за цю годину. Пікове навантаження для певної години оцінюється як високий процентиль розподілу інтенсивності надходження за цю годину (див. рис. 4.6). Таким чином, використовуючи хвіст розподілу інтенсивності надходження для передбачення пікової вимоги, можна визначити достатній обсяг ресурсів, щоб впоратися з навантаженням найгіршого випадку, якщо воно надійде. На додаток до використання спостережень за попередні дні, для прогнозування використовуються навантаження в останні кілька годин поточного дня.

Таким чином, пропонується здійснювати прогнозування навантаження на наступний інтервал часу шляхом урахування прогнозу на основі довгострокової статистики та коригуванням його за моделлю експоненційного згладжування (exponentially weighted moving average) [109], де помилки більш нових минулих періодів мають більший ваговий коефіцієнт. α (константа згладжування) – коефіцієнт, що характеризує швидкість зменшення ваг, та приймає значення від 0 до 1, чим значення цього параметра ближче до одиниці, тим більше при прогнозі враховується вплив останніх рівнів ряду. Параметри моделі встановлюються адміністратором мережі відповідно до експерименту.

Припустимо, що $\lambda_{basepred}(t)$ – передбачена інтенсивність надходження під час певного інтервалу позначеного *t*. Далі, нехай $\lambda_{obs}(j)$ – реальна інтенсивність надходження протягом інтервалу *j*. Помилка передбачення це $\left(\lambda_{obs}(j) - \lambda_{basepred}(j)\right)$. У

випадку систематичної помилки передбачення протягом останніх декількох інтервалів, прогнозоване значення протягом наступного інтервалу коригується з використанням помилки, що спостерігається:

$$\lambda_{pred}(t) = \lambda_{basepred}(t) + \alpha \sum_{j=t-h}^{t-1} (1-\alpha)^{t-1-j} \cdot \left(\lambda_{obs}(j) - \lambda_{basepred}(j)\right)^{+} \quad (4.8)$$

де h – кількість минулих інтервалів, а x^{+} позначає *max(0,x)*.

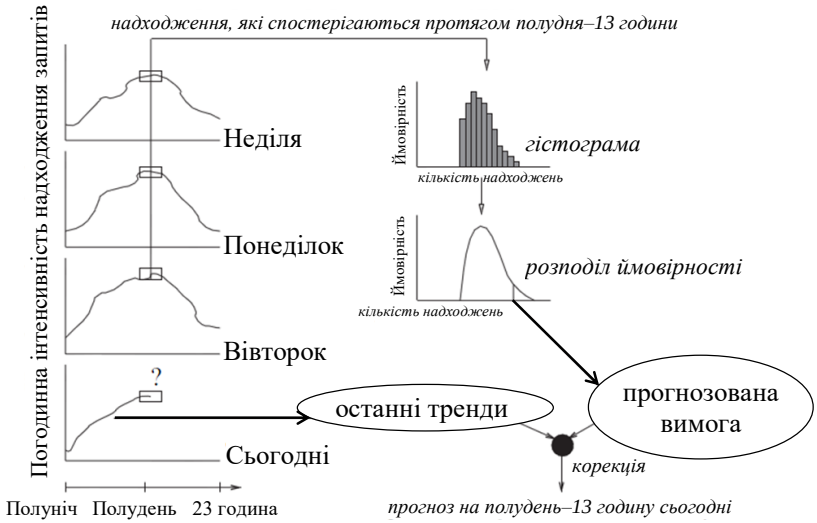


Рис. 4.6 Схема передбачення навантаження

4.7 Оцінка часу обслуговування

Двом мережевим функціям з подібними інтенсивностями надходження запитів але різними інтенсивностями обслуговування (наприклад, через різні розміри пакетів і т.д.) потрібно буде виділити різні кількості ресурсів [73].

Для оцінки інтенсивності обслуговування мережевої функції модуль прогнозування обчислює розподіл ймовірностей часу обслуговування запитів. Такий розподіл представляється гістограмою часу обслуговування запиту. По завершенні кожного запиту, ця гістограма оновлюється з урахуванням часу обслуговування цього запиту. Розподіл використовується для визначення очікуваного часу обслуговування s_i для запитів в наступному вікні адаптації. s_i може обчислюватись як середнє, медіана, або процентиль розподілу, отриманого з гістограми. Пропонується використовувати середнє значення розподілу для представлення часу обслуговування запитів мережевої функції.

4.8 Аналіз методу керування ресурсами в контексті зміни інтенсивності надходження запитів

Розглянемо задачу в системі Mathcad. Розглянемо роботу одного блоку і будемо вважати заданою базову компоненту прогнозованої інтенсивності надходження заявок, а також припускаємо наявність ресурсу одного типу.

На рис. 4.3 показано різні варіанти представлення значень часового ряду, тобто навантаження, в залежності від величини базового інтервалу, і їх вплив на точність представлення навантаження. (так помилка репрезентації для випадку інтервалів у 10 хвилин складає 7%, а для випадку 60 хвилин – вже 19%). Для порівняння на рис. 4.7 також зображена система з усього лише 4 вікнами адаптації на день – для неї помилка у представленні навантаження складає 53%, з чого можна зробити ще і той висновок, що ресурси будуть застосовуватись вкрай недоцільно, додатково на рис. 4.8 зображено систему без динамічного керування ресурсами, де помилка дорівнює 235%.

Результати моделювання системи зі змінною величиною інтервалу часу сталої конфігурації та системи без неї показали (рис. 4.9), що різниця між прогнозованим значенням та репрезентаційним може складати 9%. Якщо не застосовувати систему динамічного регулювання величини вікна керування, то відхилення складатиме 18%, тобто на 9% більше і ресурсів буде витрачатись, відповідно, більше.

Рис. 4.8 Результати моделювання системи з фіксованим керуванням

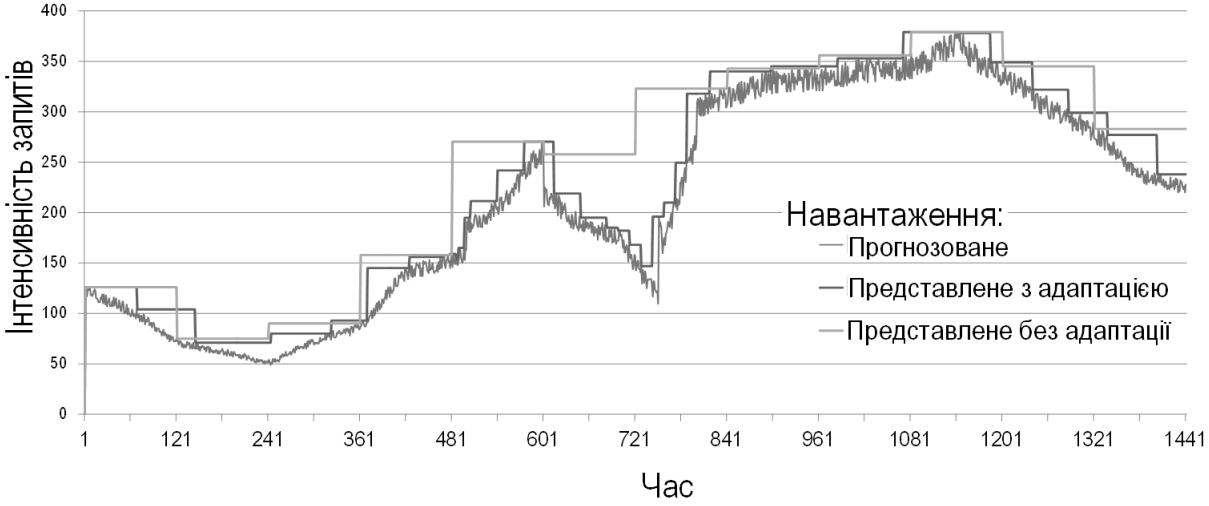


Рис. 4.9 Результати моделювання системи з динамічною зміною величини інтервалу часу сталої конфігурації та системи без неї

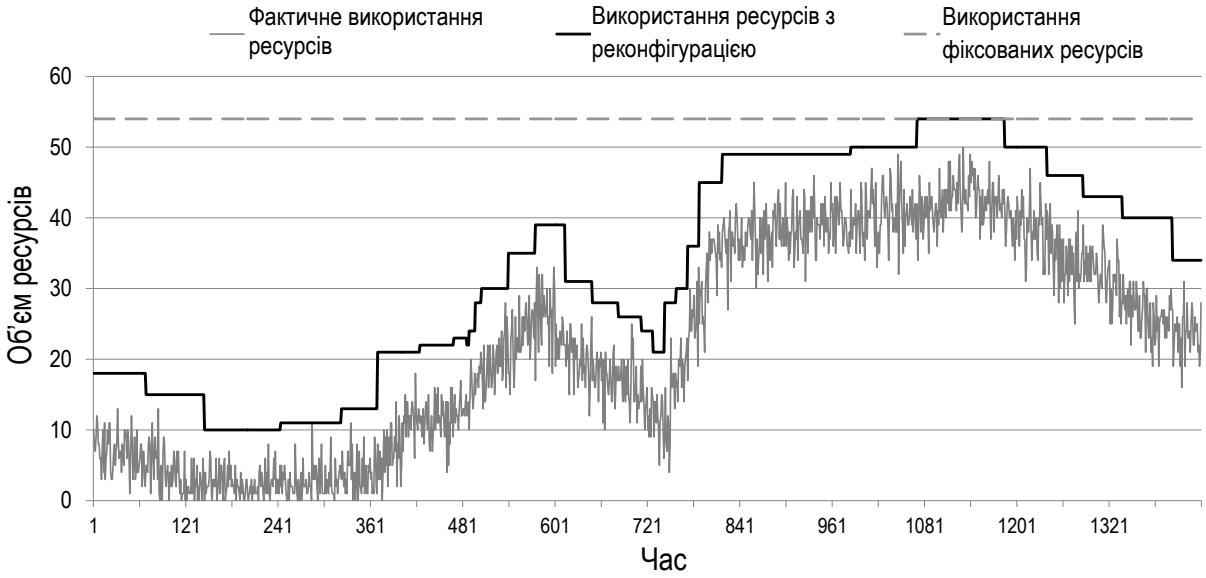


Рис. 4.10 Результат динамічного розподілу ресурсів у віртуалізованій базовій мережі

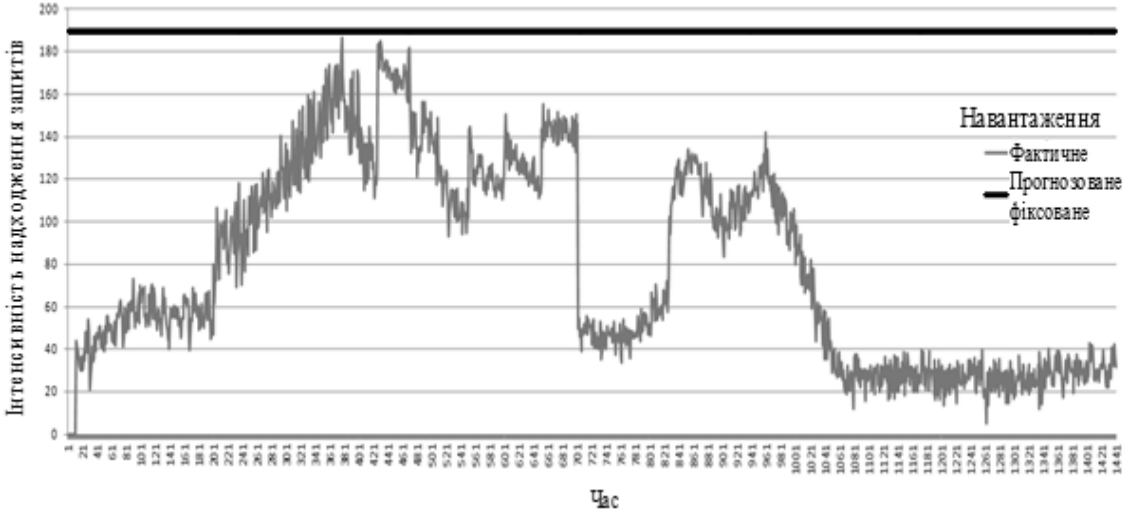


Рис. 4.8 Результати моделювання системи з фіксованим керуванням

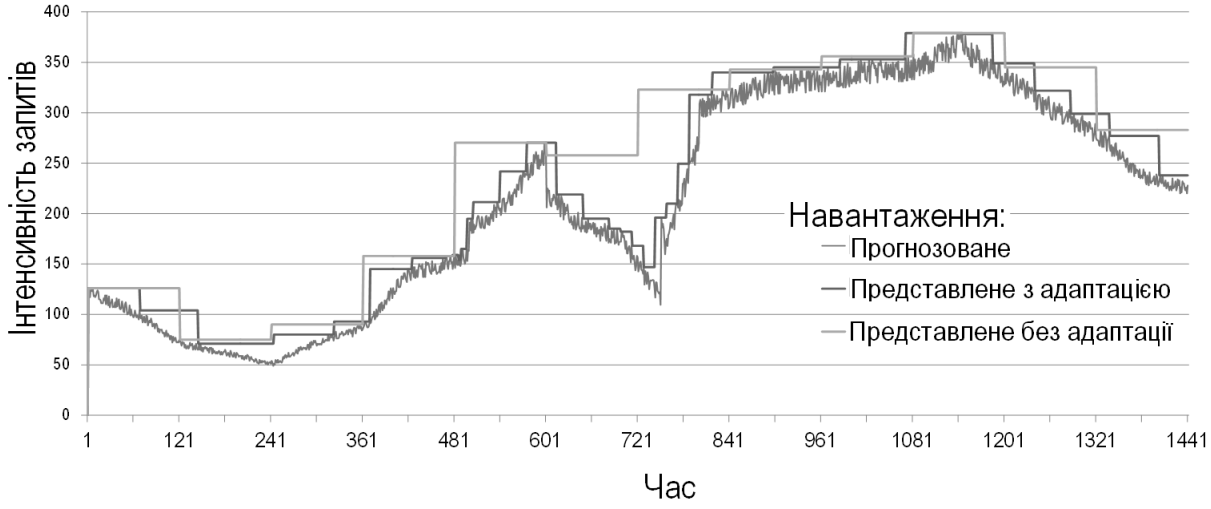


Рис. 4.9 Результати моделювання системи з динамічною зміною величини інтервалу часу сталої конфігурації та системи без неї

4.9 Експериментальне дослідження використання ресурсів у віртуалізованій базовій мобільній мережі

Експериментальне дослідження проводиться засобами імітаційного моделювання середовища Matlab. Розглядається система з 5 дата центрами. Для оцінки запропонованого підходу середня кількість вільних ресурсів в день визначалась як різниця між фіксовано виділеними ресурсами, тобто коли протягом доби завжди виділено 100% ресурсів, та динамічно виділеними ресурсами з використанням NFV. Як видно з таблиці 4.1, обєм ресурсів, що виділяються динамічно, в середньому на 42% менше, ніж у разі використання традиційного підходу розподілу. На рис. 4.10 зображено результат роботи динамічного розподілу ресурсів у віртуалізованому EPC мобільної мережі у графічному вигляді. Штрихова лінія ілюструє випадок статичного розподілу ресурсів. Сіра суцільна крива показує трафік протягом дня. Чорна крива показує обсяг ресурсів, розподілених відповідно до запропонованого методу динамічно.

Таблиця 4.1

Середній відсоток вільних ресурсів протягом дня

Дата центр	Середній % вільних ресурсів протягом дня
1	40
2	38
3	42
4	46
5	45
В середньому	42

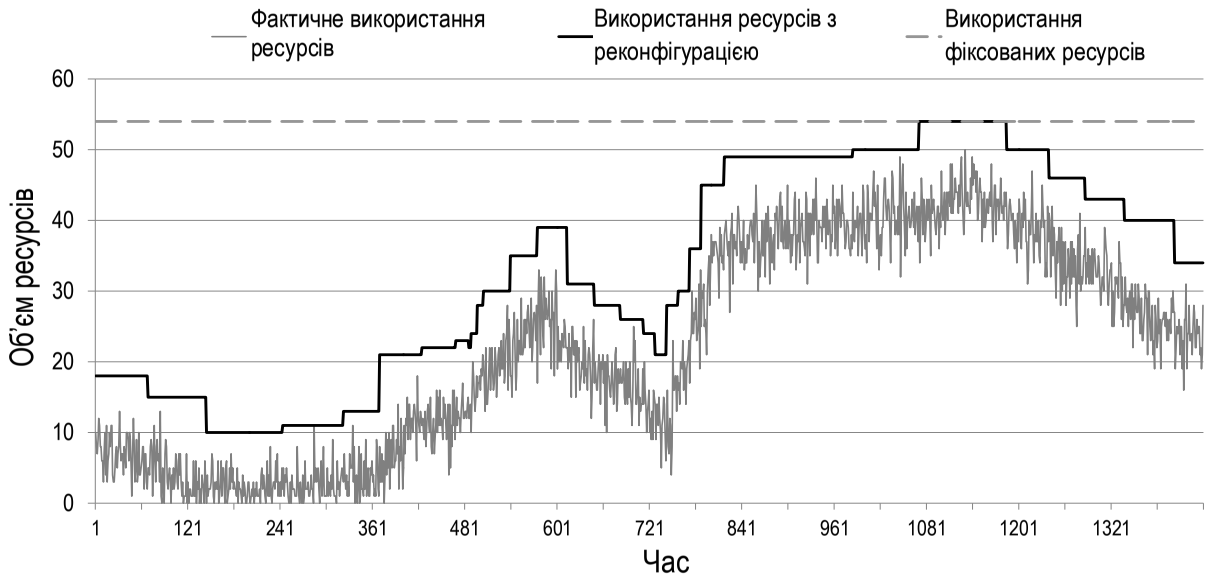


Рис. 4.10 Результат динамічного розподілу ресурсів у віртуалізованій базовій мережі

4.10 Автоматизація процесу мастабування мережевих функцій

На рис. 4.11 [51] показано обмін повідомленнями робочого процесу для операції масштабування одного VNF-екземпляра NS. Ці повідомлення були згруповані в різні етапи: збір інформації, запуск масштабування, розподіл ресурсів та звільнення ресурсів. Хоча останні два етапи можуть виконуватися незалежно, може бути, що обидва потрібні в одному сценарію масштабування (наприклад, заміна виконуваного екзмепляру VNFC іншим екземпляром більшої ємності). У цьому випадку розподіл відбувається до звільнення, щоб гарантувати безперервність сервісу (наприклад, при запуску нового екземпляру VNFC старий екземпляр можна видалити).

На етапі збору інформації NFVO збирає дані про продуктивність/несправності від VNFMs та VIMs. Показники ефективності на рівні NS/VNF повідомляються за допомогою Performance Information Available Notifications – доступних сповіщень про інформацію продуктивності, та пороговими значеннями, основаними на показниках продуктивності, за допомогою порогового повідомлення Threshold Crossed Notification (етапи 1-2). Значення індикатора VNF змінюються за допомогою EMs, а також повідомляються NFVO за участі VNFM (етап 3).

У фазі запуску масштабування (етап 4) DRPA використовує вказану вище інформацію разом із інформацією, до якої NFVO має доступ зі сховищ даних (NSD/VNFD, NS/VNF Info та стан ресурсів NFVI), щоб визначити оптимальний NS-IL. Якщо ми вважатимемо, що оптимальний NS-IL відрізняється від існуючого в VNF-IL одного екземпляра VNF, станеться процедура масштабування VNF.

NFVO вимагає, щоб VNFM масштабував екземпляр VNF, надіславши йому новий VNF-IL у повідомленні масштабування Scale VNF to Level Request.

Тепер VNFM може ініціювати операцію масштабування. З цією метою, VNFM забезпечує цю операцію управління життєвим циклом унікальним ідентифікатором, використовуючи повідомлення Scale VNF to Level Response (етап 5). VNFM використовуватиме цей ідентифікатор, щоб повідомити NFVO про початок і пізніше результат цієї операції. Нарешті, VNFM консультується з VNF Info та VNFD для порівняння поточного VNF-IL з новим. З цього порівняння можна визначити ресурси, що підлягають виділенню та/або звільненню для цієї операції масштабування.

На етапі розподілу ресурсів виділяються наступні підфази.

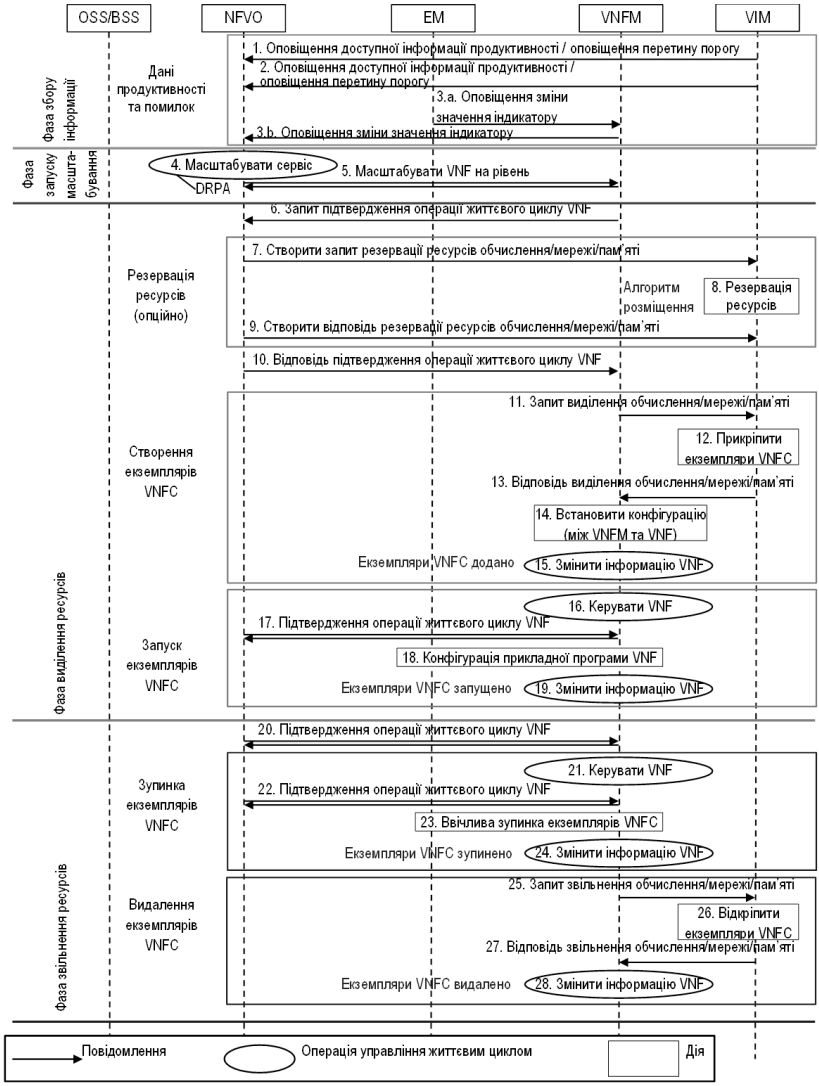


Рис. 4.11 Робочий процес процедури масштабування VNF

Резервування ресурсів (необов'язково [52]): Перед цим етапом VNFM запитує у NFVO дозвіл на видлення ресурсів. З цією метою VNFM надсилає NFVO ідентифікатори VDU та внутрішніх VL, які вказують на ресурси, які будуть виділятися (етап 6). Хоча NFVO вже мав цю інформацію після кроку 5, специфікації ETSI передбачають обмін

такою інформацією [114]. Тоді розпочинається підфаза резервування ресурсу. З результатів DRPA, NFVO знає, які NFVI-PoP повинні мати ресурси, що підлягають розподілу, а також VIM, що забезпечують доступ до цих NFVI-PoP. Тепер ці ресурси можуть бути зарезервовані для подальшого виділення. Кожен вибраний VIM отримує три запити на резервування (етап 7), по одному для кожного типу ресурсу (обчислення, зберігання та мережа), який він повинен зарезервувати в NFVI-PoP під його управлінням. Ці запити містять обмеження щодо місця розташування, що застосовуються до зазначених ресурсів. VIM використовує ці обмеження для виконання алгоритму розміщення (етап 8), приймаючи рішення щодо відповідних зон ресурсів NFVI-PoP [115], де ресурси зарезервовані. Потім VIM надсилає NFVO (етап 9) ідентифікатори зарезервованих ресурсів. Нарешті, NFVO надсилає VNFM ці ідентифікатори та інформацію про підключення [114] для кожного вибраного VIM. Тепер VNFM знає, як отримати доступ до цих VIM, і який може виділити кожен ресурс. Якщо ця підфаза не виконується, необхідно розглянути два питання. По-перше, виконуються лише кроки 6 і 10. По-друге, алгоритм розміщення тепер виконується після запиту розподілення ресурсів (див. наступну підфазу).

Створення екземплярів VNFC: VNFM надсилає ідентифікатори резервування відповідним VIM (етап 11) для виділення ресурсів (крок 12). На цьому етапі створено екземпляри VNFC та підключено їх. Ідентифікатори виділених ресурсів потім надсилаються до VNFM (етап 13). На кроці 14, VNFM запускає конфігурацію екземплярів VNFC. Нарешті, VNFM оновлює інформацію VNF у сховищі даних (етап 15), щоб відобразити створення нових екземплярів VNFC та встановлює їх стан на STOPPED.

Запуск екземплярів VNFC: Щоб запустити функціональність нових екземплярів, VNFM запускає операцію життєвого циклу Operate VNF (етап 16). Ця операція змусить (наприкінці цієї підфази) змінити стан екземплярів з STOPPED на STARTED. Коли NFVO підтверджує цю операцію (етап 17), нові екземпляри VNFC налаштовуються на рівні програми (етап 18). З цією метою VNFM спілкується з EM. Нарешті, VNFM оновлює інформацію VNF (етап 19), змінюючи стан нових екземплярів з STOPPED на STARTED. Слід зазначити, що деякі з виконуваних екземплярів VNFC можуть зазнати впливу через створення нових, і, отже, вони повинні бути (ре)конфігуровані з точки зору підключення (наприклад, нові інтерфейси, оновлені вимоги до каналу) та/або програми (наприклад, надсилання/отримання пакетів до/з нових екземплярів). У такому випадку VNFM буде вимагати від відповідних VIM та/або EM, відповідно, внесення необхідних змін. Для фази Вивільнення ресурсу виділяють дві підфази.

Зупинка екземплярів VNFC: На кроці 20 VNFM просить NFVO дозволу на звільнення ресурсів. Тоді VNFM запускає операцію життєвого циклу Operate VNF (етап 21), щоб чемно зупинити деякі екземпляри VNFC (змушуючи зупинитись екземпляр VNFC в кінці цієї підфази). На кроці 23, затронуті екземпляри (ре)конфгуруються (слідуючи відповідно до стратегій, що зазначені у підфазі Запуску екземплярів VNFC), а екземпляри, що підлягають завершенню, вимикаються. Нарешті, стан зупинених екземплярів змінюється з STARTED на STOPPED (етап 24).

Видалення екземплярів VNFC: VNFM надсилає відповідним VIM ідентифікатори ресурсів, які розміщують та з'єднують зупинені екземпляри VNFC (етап 25). На даному етапі ці екземпляри видаляються (етап 26). Потім VIM надсилають VNFM

ідентифікатори звільнених ресурсів (етап 27). Після отримання цих ідентифікаторів VNFM оновлює інформацію VNF (етап 28), щоб відобразити видалення екземпляра.

Висновки

1. Коректне керування обчислювальними ресурсами в дата центрах для задоволення вимог до якості обслуговування при цьому мінімізуючи споживання ресурсів є складним завданням.

2. У присутності динамічно змінюваних навантажень потрібні методи динамічного керування конфігурацією ресурсів, для того щоб забезпечити гарантії для віртуалізованих мережевих функцій, що працюють на загальних дата центрах. Для вирішення цієї проблеми використовуємо системну архітектуру, яка поєднує в собі онлайн вимірювання з методами прогнозування і розподілу ресурсів.

3. Розроблено метод визначення розміру інтервалу часу сталої конфігурації ресурсів для точного визначення шаблонів навантаження, який передбачає змінну величину інтервалів з метою оптимального використання ресурсів та забезпечення заданої якості обслуговування.

4. Запропоновано метод, який поєднує прогностичне і реактивне надання ресурсів: прогностичне надання ресурсів виділяє ресурси завчасно в очікуванні певного пікового навантаження, в той час як реактивне надання ресурсів вносить корективи після того, як спостерігалося аномальне збільшення навантаження. Іншими словами, у той час як прогностичне надання ресурсів намагається «випереджати» коливання навантаження, реактивне надання ресурсів дозволяє хостинговій платформі бути гнучкою до відхилень від очікуваного навантаження.

Контрольні запитання

1) Опишіть постановку задачі керування ресурсами мережевих функцій.
2) Опишіть систему динамічного керування ресурсами.
3) Які доступні механізми моніторингу надаються на різних рівнях?
4) Як визначається інтервал сталої конфігурації ресурсів?
5) Як визначається інтенсивність обслуговування запиту?
6) Як виглядає схема передбачення навантаження?
7) Опишіть метод прогнозування навантаження.
8) Як виглядає схема передбачення навантаження?
9) Опишіть метод прогнозування навантаження.
10) Який виграш системи з динамічною зміною величини інтервалу часу сталої конфігурації у порівнянні з системою без неї?
11) Який виграш у динамічному використанні ресурсів у віртуалізованій базовій мобільній мережі?
12) Як відбувається процес масштабування VNF?

РОЗДІЛ 5
РЕКОНФІГУРАЦІЯ МЕРЕЖІ ПІСЛЯ ПЕРЕВАНТАЖЕННЯ АБО ЗБОЮ

5.1 Модель мережі та постановка задачі

Фізична мережа задана у вигляді графа $SN=(N,NE)$, де N є множиною фізичних вузлів і NE – множиною каналів. Кожен канал $(n_1,n_2)\in NE$, $n_1,n_2\in N$ має максимальну пропускну здатність $c(n_1,n_2)$ і мережеву затримку $L(n_1,n_2)$, а кожен вузол $n \in N$ пов'язаний з певними ресурсами c_n^i, $i\in R$, де R – множина типів ресурсів. Мережа зв'язку представлена множиною сервісів (або запитів віртуальної мережі) K, які вбудовуються в фізичну мережу. Запит віртуальної мережі k, $k\in K$, можна представити як зважений граф $G_k=(V_k,E_k)$, де V_k є множиною віртуальних вузлів, що містить h_k елементів і позначається як $V_k=(v_{k,1},v_{k,2},...,v_{k,hk})$, де $v_{k,j}$ означає j-у мережеву функцію у ланцюзі функцій k E_k є множиною віртуальних каналів $e_k(v_{k,j},v_{k,g})\in E_k$. Вимоги смуги пропускання каналу між двома функціями, $j1$ і $j2$, що відносяться до сервісу $k\in K$ позначається як $d_k^{(j1,j2)}$, $d_k^{j,i}$ – кількість ресурсу типу i, що виділяється для мережевої функції j сервісу k. Булеві змінні $x_n^{k,j}$ вказують, чи мережева функція j, пов'язана з $k \in K$, розташовується на фізичному вузлі n, змінні $f_{(n1,n2)}^{k,(j1,j2)}$ визначають, чи фізичний канал $(n1,n2)$ використовується у шляху між $j1$ та $j2$ для запиту k. L_k – максимальна затримка для запиту k. $costN(i,n)$ – вартість зайнятої одиниці ресурсу i на фізичному вузлі n, і $costL(n_1,n_2)$ – вартість зайнятої одиниці пропускної здатності на фізичному каналі $(n_1,n_2) \in NE$. $suit_n^{k,j}$ означає що функція j з запиту k може бути розміщена на вузлі n.

MN являє собою множину вузлів керування, де $MN \subseteq N$, які відповідають за функціонування пропонованого механізму реконфігурації після відмови або перевантаження. Кожен керуючий вузол пов'язаний з одним або декількома вузлами фізичної мережі і виконує кроки, необхідні для відновлення належного функціонування мережі. Процес призначення вузлів керування і критерії, які враховуються при виборі вузлів керування, будуть досліджені далі.

У запропонованому підході припускаємо, що відображення запитів віртуальної мережі вже виконано і вузли керування у фізичній мережі призначаються, як описано нижче.

Процес відображення віртуальної мережі відбувається в два етапи [93]: відображення вузлів ($M_N : V_k \rightarrow N$) і відображення каналів ($M_L : E_k \rightarrow NE$).

Проблема, що розглядається далі, полягає в тому, як перемістити розміщені на вузлі з відмовою віртуальні вузли з метою мінімізації витрат відновлення вузла відмови і періоду переривання сервісу. У пропонованому підході підкреслюється, що будь-яке переміщення віртуального вузла повинно виконуватись локально і бути координованим тільки вузлами керування.

5.2 Оптимальне розміщення вузлів керування у мережах, заснованих на NFV

Припускаємо, що вузли керування (далі – менеджери) можуть розміщуватись в вузлах *N*.

При заданій кількості менеджерів *A* існує скінченна множина з $\binom{|N|}{A}$ можливих розташувань, відповідно, задача розміщення менеджерів є задачею багатокритеріальної комбінаторної оптимізації. Метою задачі є знаходження таких розташувань менеджерів з множини можливих розташувань розміру $A - Pl_A = \{Pl \in 2^{|N|} \big| |Pl| = A\}$, що є оптимальними відповідно до деякої цільової функції.

Метою оптимізації є визначення місця розташування кожного менеджеру при заданій їх кількості *A*, так що мінімізується функція загальних витрат $Cost_A(\{p_n{:}n \in N\})$, де p_n – булева змінна, яка рівна одиниці, якщо менеджер розміщується в точці *n*. Задача оптимізації буде мати вигляд (51).

$$\min_{\{p_n:n\in N\}} U_A$$
$$\text{при обмеженні } \sum_n p_n = A \qquad (5.1)$$

Основною метою хорошого розміщення менеджерів є мінімізація затримок між вузлами і менеджерами в мережі. Проте, розглядати тільки затримки не достатньо. Розміщення менеджерів повинно також враховувати певні обмеження стійкості.

На рис. 5.1 показані різні питання, які необхідно враховувати при оцінці стійкості розміщення. Нижче коротко пояснимо ці питання, і що необхідно, щоб бути стійким по відношенню до них. На рис. 5.1 показано нормалізовані затримки між вузлами та інтенсивність навантаження на вузлах.

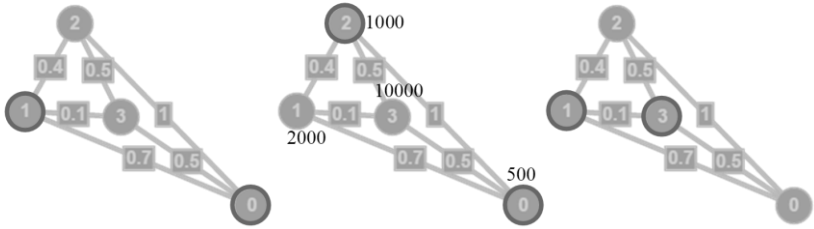


Рис. 5.1 Розміщення по менеджерів за різним критеріям: (а) мінімальної затримки до менеджера; (б) мінімального дисбалансу навантаження на менеджерів; (в) мінімальної затримки між менеджерами

Аналогічно до [71], припустимо, що вузли призначаються до їх найближчого менеджера, використовуючи в якості метрики затримки $dl_{g1,g2}$ між вузлом *g1* і менеджером *g2*. Кількість вузлів на менеджера може бути незбалансованою – чим більше вузлів менеджер повинен контролювати, тим вище навантаження на цього менеджера. Якщо кількість запитів вузла до менеджера в мережі збільшується, аналогічно поводить себе і ймовірність додаткових затримок через черги в системі керування. Для того, щоб бути стійким від перевантаження менеджера, призначення вузлів різним менеджерам повинно бути добре збалансованим.

Цілком очевидно, що одного менеджера не досить, щоб досягти стійкості в мережі. Проте, коли кілька менеджерів розміщуються в мережі, логіка керування мережі розподіляється по декількох менеджерах і ці менеджери повинні синхронізуватися, щоб

підтримувати несуперечний глобальний стан. Залежно від частоти синхронізації між менеджерами, затримка між окремими менеджерами грає важливу роль.

5.2.1 Затримка вузол-менеджер

Аналогічно, як представлено в [70], на основі матриці *dl*, що містить відстані найкоротших шляхів між усіма вузлами, максимальний час затримки передачі між вузлом і менеджером для певного розміщення менеджерів може бути визначений як

$$U_A^{latency}(p)=max\ ddc_n, \quad (5.2)$$

де ddc_n – максимальна затримка передачі від вузла мережі до менеджеру в точці n; ddc_n розраховується наступним чином:

$$ddc_n = \max_{g \in N} latency_g \cdot \pi_{g,n},$$

де $latency_g$ – затримка між менеджером та вузлом g, $latency_g = \min_{\{n: n \in N \cap p_n = 1\}} dl_{g,n}$;

$\pi_{g,n}$ – булева змінна, яка рівна одиниці, якщо вузол g обслуговується менеджером розміщеним в точці n.

Розглядаємо не середнє, але максимальне значення затримки, так як середнє приховує значення найгіршого випадку, які є важливими, коли стійкість повинна бути поліпшена.

5.2.2 Збалансований розподіл навантаження менеджерів

Залежно від ситуації, може бути бажано мати приблизно рівне навантаження на всіх менеджерів, так що жоден менеджер не перевантажується, а у інших мало роботи. Далі розглядаємо збалансований розподіл вузлів між менеджерами. Як формальну метрику вводимо баланс розміщення або, вірніше, дисбаланс, $U_A^{imbalance}$, тобто відхилення від повністю збалансованого розподілу, як різниця між навантаженням на найбільш завантаженому менеджері і найменш завантаженому менеджері.

$U_A^{imbalance}$ визначається в такий спосіб:

$$U_A^{imbalance}(p)=max\ ldc_n - min\ ldc_n, \text{ де } ldc_n > 0, \quad (5.3)$$

ldc_n – навантаження на менеджер в точці n; ldc_n розраховується наступним чином:

$ldc_n = \sum_{g \in N} load_g \cdot \pi_{g,n}$,

де $load_g$ – коефіцієнт навантаження на вузол g.

5.2.3 Затримка менеджер-менеджер

Як останній аспект стійкого розміщення менеджерів, розглянемо як затримка між менеджерами може враховуватися при виборі розміщення менеджерів. Формально, затримка між менеджерами $U_A^{interlatency}$ визначається як найбільша затримка між будь-якими двома менеджерами при заданому розміщенні:

$$U_A^{interlatency}(p) = \max_{\{n,g: n,g \in N \cap p_n = 1, p_g = 1\}} dl_{g,n}. \quad (5.4)$$

Загалом, розміщення з урахуванням затримки між менеджерами мають тенденцію розміщувати всіх менеджерів набагато ближче один до одного. Це збільшує максимальну затримку від вузлів до менеджерів.

5.2.4 Цільова функція оптимізації

Таким чином,

$$U_A = wu^{latency} \times U_A{}^{latency}(p) + wu^{imbalance} \times U_A{}^{imbalane}(p) + wu^{interlatency} \times U_A{}^{interlatency}(p), \quad (5.5)$$

де *wu* – множина вагових коефіцієнтів.

На рис. 5.2 показані можливі рішення по розміщенню двох менеджерів у мережі з 10 вузлами. Оптимальне значення показує, що усі оптимізаційні цілі не можуть бути досягнуті одночасно.

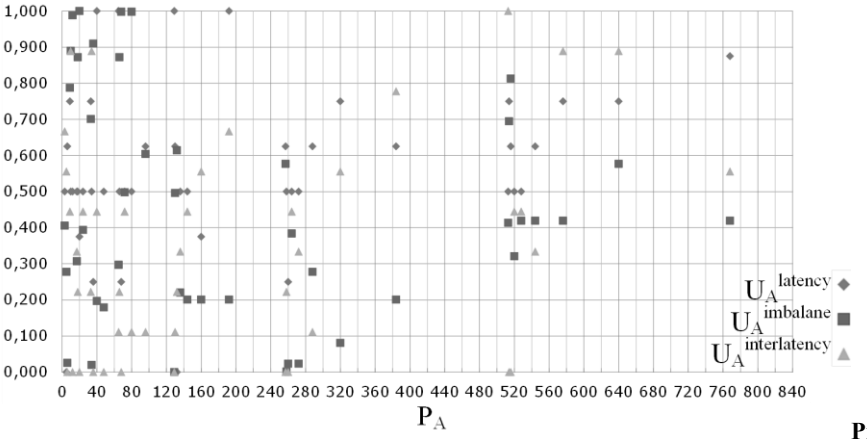


Рис. 5.2 Простір рішень оптимізаційної задачі розміщення менеджерів

5.2.5 Конфігурація вузлів менеджерів

Кожен менеджер (наприклад, вузол 3 на рис. 5.3 управляє вузлом 2) підтримує таблицю (наприклад, таблиця 5.1), що містить ідентифікатори віртуальних мереж, які мають віртуальні вузли розміщені на керованому фізичному вузлі [96]. Крім того, таблиця містить посилання на розміщені на фізичних вузлах віртуальні вузли (наприклад, вузли c та d на рис. 5.3) і суміжні до них віртуальні вузли (наприклад, вузли a та b на рис. 5.3). Ця таблиця безперервно оновлюється (наприклад, кожен раз, коли приймається або відхиляється запит віртуальної мережі).

Таблиця 5.1

Таблиця на менеджері 3 на рис. 5.3

ID віртуальної мережі	Віртуальний вузол	Ємність	Прилеглі вузли
1	c	10	a на вузлі 0 b на вузлі 1
2	d	15	f на вузлі 4 e на вузлі 5

На рис. 5.3 показаний приклад фізичної мережі з двома вже вбудованими віртуальними мережами. Вузол з відмовою (вузол 2) був призначений менеджеру (вузлу 3), який відповідає за переміщення віртуальних вузлів.

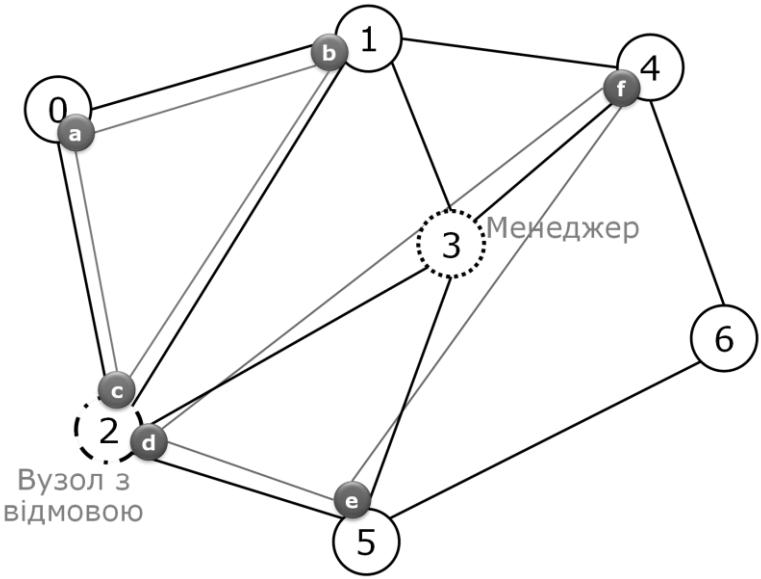


Рис. 5.3 Фізична мережа з вузлом з відмовою, на якому розміщується два віртуальних вузла

5.3 Алгоритм відновлення мережі після відмови

Процес переміщення вузлів віртуальної мережі, розміщених на вузлі, який відмовив, $v_{k,j}^{fail}$, запускається, коли система відправляє запит на відновлення відповідному вузлу-менеджеру (наприклад, вузлу 3 на рис. 5.3). Процес відновлення для кожного порушеного вузла віртуальної мережі протікає в такий спосіб: менеджер направляє запит на відновлення до всіх вузлів фізичної мережі, на яких розміщаються віртуальні вузли, суміжні з ураженими віртуальними вузлами (наприклад, вузол 0 і вузол 1 на рис. 5.3). Кожен з цих вузлів будує дерево найкоротших шляхів (Shortest Path Tree – SPT) до всіх вузлів фізичної мережі на відстані не більше l (поріг встановлюється провайдером послуг), де коренем SPT виступає сам цей вузол. Менеджер використовує ці шляхи, щоб вибрати вузол з оптимальною відстанню до всіх вузлів фізичної мережі, де розташовані вузли віртуальної мережі прилеглі до несправного вузла. Цей вузол в кінцевому рахунку стає оптимальним кандидатом для розміщення віртуального вузла з відмовою. Крім того, ємність кінцевих вузлів шляхів з SPT повинна бути не менше ємності віртуального вузла, розміщеного на вузлі з відмовою (наприклад, вузла c). Обираємо вузол з мінімальною вартістю шляху до всіх кореневих вузлів у деревах SPT та мінімальною вартістю

обчислень. Рис. 5.4 містить опис псевдокоду алгоритму відновлення вузла після відмови і виконується для всіх $\{v_{k,j} : x_n^{k,j}=1 \ \& \ n = failed\}$.

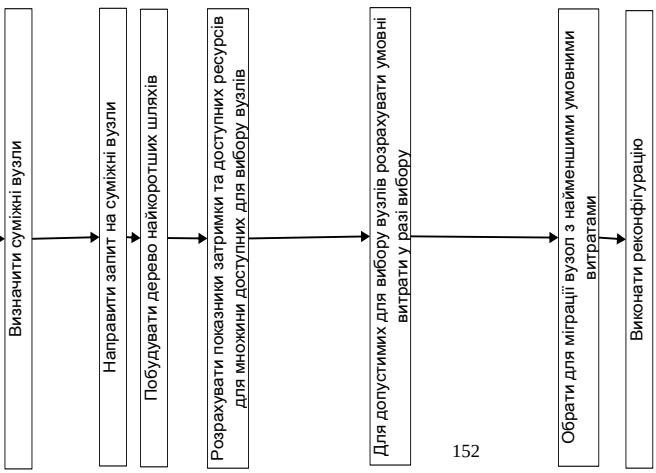


Рис. 5.4 Алгоритм відновлення вузла з відмовою

5.4 Алгоритм відновлення у разі перевантаження вузла

У мережі існує ймовірність відмови вузла через перевантаження. Для виконання відновлення при відмові вузла в перевантаженій мережі, виконується процедура реконфігурації для міграції віртуальних вузлів, розміщених на перевантаженому фізичному вузлі (рис. 5.5).

Процес відновлення починається з сортування всіх віртуальних вузлів, розміщених на перевантаженому фізичному вузлі. Критерієм (CRT в Алгоритмі рис. 5.6), що використовуються для сортування цих вузлів віртуальної мережі, є ємність віртуальних вузлів. Потім виконується процедура відновлення на першому відсортованому вузлі віртуальної мережі, що має ємність рівну перевантаженню, для переміщення на новий вузол фізичної мережі.

Рис. 5.5 Схема відновлення вузла з перевантаженням

n=перевантажений

S_1←Відсортувати віртуальні вузли, що розміщуються на n у зростаючому порядку на основі критерію CRT

Вибрати з S_1 перший віртуальний вузол $v_{k,j}$,ємність ресурсів не менше перевантаженої ємності $d_k{}^{j,i} \geq \Delta c_n{}^{i} \ \forall i \in R$

Виконати алгоритм Відновлення Вузла

Рис. 5.6 Алгоритм відновлення вузла з перевантаженням

5.5 Реконфігурація перевантажених мереж

Коли навантаження або ресурси змінюються, деякі VNF можливо доведеться перемістити. Існує ймовірність того, що знайти новий вузол-кандидат для вузла віртуальної мережі, розміщеного на вузлі з відмовою, не вийде. В такому випадку виконується процедура реконфігурації для міграції одного або декількох віртуальних вузлів для переміщення розміщених вузлів віртуальної мережі. Розглянемо задачу міграції як задачу оптимізації, яка спрямована на мінімізацію загальної вартості міграції при обмеженнях допустимої затримки і обчислювальних ресурсів.

Метою оптимізації є знаходження розташування ланцюгів сервісів мережі (тобто розміщення мережевих функцій та розподіл ресурсів), так щоб мінімізувати витрати на зайняті ресурси каналів і вузлів у фізичній мережі, при цьому задовільняючи вимоги трафіку. Сформулюємо цільову функцію у вигляді лінійної комбінації (з ваговими коефіцієнтами *a*, *b*, *c*, *e*) вартісних виразів.

Визначимо бінарну змінну $x_n^{k,j} \in \{0,1\}$, для позначення того, що VNF *j* пов'язана з ланцюгом сервісу *k* розміщується на вузлі *n* після міграції. Індикатор $x_n^{k,j}=0$ означає, що VNF *j* не розміщується на вузлі *n* після міграції; в іншому випадку *j* розміщується на вузлі *n* після міграції.

Введемо бінарну змінну $y_n^{k,j}$ для індикації стану мережі перед міграцією. Вона схожа зі змінною $x_n^{k,j}$, $y_n^{k,j}=0$ означає, що VNF *j* сервісу *k* не перебуває на вузлі *n* до міграції; в іншому випадку, *j* розташована на вузлі *n* до міграції.

Таким чином, можемо використовувати індикатор $I^{k,j}$, щоб вказати чи VNF *j* сервісу *k* було переміщено в поточному процесі міграції.

$$I^{k,j} = \sum_{n \in N} x_n{}^{k,j} \cdot y_n{}^{k,j}$$

$I^{k,j} = 0$ вказує, що VNF було переміщено в поточному процесі міграції, і $I^{k,j} = 1$ вказує, що VNF не було переміщено.

x_n позначає чи *n*-ий фізичний сервер працює або ні після міграції.

$$x_n = \begin{cases} 1 \text{ (сервер } n \text{ працює)} \\ 0 \text{ (в іншому випадку)} \end{cases}$$

y_n позначає чи *n*-ий фізичний сервер працює або ні перед міграцією.

$$y_n = \begin{cases} 1 \text{ (сервер } n \text{ працює)} \\ 0 \text{ (в іншому випадку)} \end{cases}$$

Для того щоб розглянути ресурси, які споживаються при міграції та запуску серверів, вводимо такі вирази:

- B_n позначає необхідні витрати b_n для запуску *n*-ого серверу. $B_n = b_n x_n (x_n - y_n)$;
- $L_i^{k,j}(n \to n')$ позначає використання ресурсу *i* для міграції VNF *j* з ланцюгу сервісу *k* з серверу *n* на сервер *n'*.

$L_i^{k,j}(n \to n') = l_i(d^{k,j}) + l'_i(d^{k,j})$,

де $l_i(x)$ – функція використання ресурсу *i* для міграції з серверу,

$l'_i(x)$ – функція використання ресурсу *i* для міграції на сервер.

Цільова функція буде визначатися як:

$$MCost = a \cdot \sum_{n \in N} \left(B_n + x_n \cdot cost(n)\right) +$$
$$+ b \cdot \sum_{n \in N} \sum_{k \in K} \sum_{j \in V} \sum_{i \in R} x_n{}^{k,j} \cdot d_i{}^{k,j} \cdot costN(i,n) +$$
$$+ c \cdot \sum_{(n_1,n_2) \in NE} costL(n_1,n_2) \cdot \sum_{k \in K} \sum_{(j_1,j_2) \in E} f_{(n_1,n_2)}{}^{k,(j_1,j_2)} \cdot d_k{}^{(j_1,j_2)} +$$
$$+ e \cdot \sum_{n \in N} \sum_{n' \in N} L_i{}^{k,j}(n \to n') x_{n'} (x_{n'} - y_n) \quad (5.6)$$

Використовуючи наведені вище міркування, формулюємо задачу наступним чином.

Цільова функція:

$$Min\ MCost$$

З обмеженнями:

$$\sum_{n \in N} x_n{}^{k,j} = 1 \ \ \forall k \in K, j \in V, \quad (5.7)$$

$$x_n{}^{k,j} \le suit_n{}^{k,j} \ \ \forall k \in K, j \in V, n \in N, \tag{5.8}$$

$$\sum_{k\in K}\sum_{j\in V} \begin{matrix} x_n{}^{k,j} \cdot d_i{}^{k,j} + y_n{}^{k,j} \cdot \left(1 - I^{k,j}\right) \cdot l_i\left(d^{k,j}\right) + \\ + x_n{}^{k,j} \cdot (1 - I^{k,j}) \cdot l'_i\left(d^{k,j}\right) \end{matrix} \le c_n{}^i \ \ \forall n \in N, i \in R, \tag{5.9}$$

$$\sum_{t\in K}\sum_{(j_1,j_2)\in E} f_{(n_1,n_2)}{}^{k,(j_1,j_2)} \cdot d_k{}^{(j_1,j_2)} \le c(n_1,n_2) \ \ \forall (n_1,n_2) \in NE, \tag{5.10}$$

$$\sum_{(n,w)\in L} f_{(w,n)}{}^{k,(j_1,j_2)} - f_{(n,w)}{}^{k,(j_1,j_2)} = x_n{}^{k,j_1} - x_n{}^{k,j_2} \ \ \forall k \in K, n \in N, (j_1,j_2) \in E, \tag{5.11}$$

$$x_n{}^{k,j}, f_{(n_1,n_2)}{}^{k,(j_1,j_2)} \in \{0,1\} \ \ \forall k \in K, j \in V, n \in N, (j_1,j_2) \in E, (n_1,n_2) \in NE, \tag{5.12}$$

$$\sum_{(j_1,j_2)\in E}\sum_{(n_1,n_2)\in NE} f_{(n_1,n_2)}{}^{k,(j_1,j_2)} \cdot L(n_1,n_2) \le L_k \ \forall k \in K, \tag{5.13}$$

$$\sum_{n\in N} x_n{}^{k,j} \sum_{i\in R}\left(\frac{1}{\frac{d_k{}^{j,i}}{s_{n,i}{}^{k,j}} - \lambda^{k,j}}\right) \le P_k{}^j \ \ \forall k \in K, j \in V \tag{5.14}$$

Отже, цільова функція (5.6) представляє собою лінійну комбінацію чотирьох вартісних виразів та направлена на мінімізацію: вартості запуску та використання серверу, використання ресурсів серверу, використання каналів зв'язку та використання ресурсів для міграції. Обмеження (5.7) гарантує однократність розміщення мережевих функцій, а (5.8) – адміністративну можливість розміщееня на вузлі. Вирази (5.9) та (5.10) являють собою обмеження на ресурси фізичних вузлів і каналів, тобто забезпечують той факт, що кількість задіяних на вузлі ресурсів не перевищує кількості доступних. Вираз (5.11) представляє собою обмеження щодо збереження потоку для всіх шляхів у фізичній мережі, тобто що вхідний потік на вузлі дорівнює вихідному потоку. Вираз (5.12) гарантує, що змінні у задачі є булевими. Вирази (5.13) та (5.14) представляють собою обмеження на час передачі телекомунікаційними каналами та час обробки вузлами обслуговування відповідно, і забезпечують дотримання заданих часових вимог до обслуговування сервісу.

5.6 Оцінка методу локальної реконфігурації мережі

За результатами моделювання (рис. 5.7) запропонований метод показав до 27% менші умовні витрати у порівнянні зі стратегією направленою на мінімізацію затримки, при цьому затримка знаходилась у допустимих межах але була на 20% більшою.

Рис. 5.7 Результати відновлення вузла

Висновки

1. Розглянуто метод реконфігурації системи у випадку відмови вузла через перевантаження або збій, який спирається на співпрацю множини розподілених керуючих вузлів – менеджерів, розміщених на ряді фізичних вузлів. У дослідженні вирішується проблема розташування локальних вузлів керування з урахуванням параметрів затримки та стійкості, пропонується цільова функція комбінаторної оптимізаційної задачі розташування вузлів керування.

2. Пропонована схема відновлення мережі зменшує вартість відновлення вузла після збоїв, зазнавану затримку і час переривання сервісу під час цього процесу, максимізуючи дохід телекомунікаційного провайдера.

Контрольні запитання

1) Сформулюйте постановку задачі реконфігурації мережі після перевантаження або збою.
2) Які параметри необхідно враховувати при оцінці стійкості розміщення вузлів керування у мережі?
3) Як враховується затримка вузол-менеджер?
4) Як враховується розподіл навантаження менеджерів?
5) Як враховується затримка менеджер-менеджер?
6) Як виглядає цільова функція оптимізації?
7) Як виглядає таблиця вузлів на менеджері?
8) Який алгоритм відновлення вузла з відмовою?
9) Який алгоритм відновлення вузла з перевантаженням?
10) Як проходить реконфігурація перевантажених мереж?

РОЗДІЛ 6 МОДЕЛІ КОНТРОЛЮ ЯКОСТІ ОБСЛУГОВУВАННЯ КІНЦЕВИХ КОРИСТУВАЧІВ

6.1 Моделі контролю якості обслуговування

Забезпечення якості обслуговування для телекомунікаційних систем включає в себе методи контролю та керування роботою телекомунікаційної системи. На сьогоднішній день розроблено безліч механізмів адаптивного реагування на зниження якості обслуговування. Можна виділити наступні групи механізмів:

1. Контроль показників обслуговування черг у вузлах зв'язку.
2. Контроль завантаженности мережі.
3. Управління організацією подій для диференційованого обслуговування мультисервісних потоків
4. Контроль параметрів потоків абонентських даних
5. Методи зворотного зв'язку - механізми попередження джерел даних про можливі перевантаження вибраного напрямку зв'язку.
6. Методи інжинірингу трафіку для планування рівномірного завантаження ресурсів мережі.

Робота системи контролю якості обслуговування для телекомунікаційної системи з віртуалізацією мережевих функцій пов’зана з контролем та керуванням телекомунікаційної системи, а також з контролем роботи інформаційно-обчислювальної системи. Інформаційно-обчислювальна система забезпечення мобільного зв’язку, розгортається на базі групи дата центрів територіально розподілених та обєднаних у єдиний обчислювальний простір за технологією гібридна хмарна система.

Наразі досліджуються два основні підходи до віртуалізації:

- Віртуалізація функцій керування мережею, що передбачає виконання завдань пов’язаних з обслуговуванням службових потоків у відокремлених обчислювальних системах розміщених у дата центрах із гетерогенною хмарною інфраструктурою. Всі потоки даних проходять через телекомунікаційні системи, їх не перенаправляють до хмар, відповідно не створюють надлишкове навантаження на телекомунікаційну систему.
- Віртуалізація всіх функцій телекомунікаційної системи, коли для обслуговування телекомунікаційного сервісу його направляють до дата центру, приклад віртуальна базова станція з віртуалізацією функції кодування/декодування, яка передбачає виконання операцій не лише з потоками керування, але й з потоками даних.

Контроль якості обслуговування передбачає перевірку відповідності параметрів обслуговування заданим пороговим значенням показників якості обслуговуваня на кожному етапі обслуговування сервісу. Порогові значення параметрів якості обслуговування користувачів зберігаються у базах підсистеми PCC (Policy and Charging Control), контроль параметрів обслуговування здійснюється для кожного користувача індивідуально відповідно до договору на обслуговування. Основними класичними, в термінах теорії масового обслуговування, параметрами QoS (Quality of service) є затримка у вузлі обслуговування, втрати заявок через перевищення часу очікування на

обслуговування, втрати заявок через перевантаження черг на обслуговуючих пристроях. Саме ці параметри досліджують.

На сьогоднішній день розроблено групу стандартів та специфікацій [116, 117], в яких описано підходи до контролю якості обслуговування у системах мобільного зв'язку 3G, 4G. Основною підсистемою яка перевіряє відповідність поточних показників якості заявленим у PCC показникам є PCRF (Policy and Charging Rules Function). В стандартах описані основні функції та процедури забезпечення контролю та керування якістю, а також взаємодія підсистем мобільного зв'язку.

Впровадження підходів віртуалізації змінить функціонування системи забезпечення якості обслуговування.

При реалізації підходу з віртуалізацією системи керування (Рис. 6.1) основною задачею є забезпечення вимог до якості функціонування системи керування, а саме своєчасність прийняття керівних рішень. Моніторинг роботи системи мобільного зв'язку здійснює підсистема SGSN. При виявленні невідповідності показників доступності послуг пороговим значенням, виникає потреба у оптимізації роботи системи керування. Ефективність функціонування віртуалізованої системи керування залежить від ефективності організації обчислювального процесу.

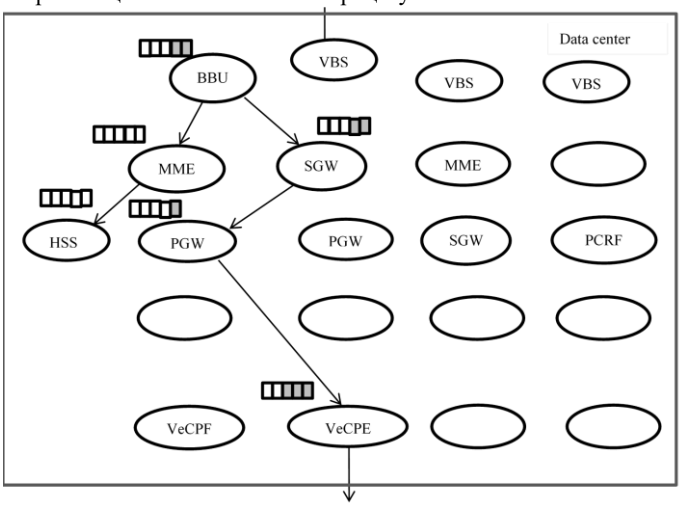


Рис. 6.1 Обслуговування потоків в Інфраструктурі NFV

Розроблено також архітектурну модель логічної структури віртуальної мережі, на базі якої розгортається інфраструктура NFV. Кількість вузлів у віртуальній мережі а також кількість віртуалізованих сутностей залежить від структури вхідних потоків від абонентів та розташування точок їх входу (надходження) до системи обслуговування. *Особливістю віртуалізованого середовища* є гнучкість та можливість динамічної зміни структури мережі, місце обслуговування потоків, та кількість NFV-ресурсів, залучених

до процесу обслуговування. Основним засобом контролю якості обслуговування гібридних сервісів є контроль показників обслуговування.

6.1.1 Вимоги до якості в 5G мережі

В процесі еволюції механізму управління QoS в GSM/UMTS/LTE мережах сталася міграція управління QoS з рівня користувацького обладнання до керування на рівні мережі. Цей підхід також збережеться в мережах 5G.

Механізми керування QoS в мережах 5G повинні забезпечити пріоритет відео і VoIP трафіку над іншими сервісами. Сервіс потокового відео без буферизації дуже чутливий до затримок в мережі, тому одним з найбільш важливих параметрів, який визначає вимоги до QoS є загальний час затримки пакетів (packet delay budget – PDB). У Таблиця 6.1 наведені вимоги до затримки в 3G/4G/5G мережах.

Таблиця 6.1

Вимоги до затримки в 3G/4G/5G мережах

Терміни QoS	Запланована затримка пакетів (мс)		
	3G	4G	5G
Без гарантування якості	Не визначено	100-300	Не визначено
З гарантованою якістю	100-280	50-300	1

Ці дані показують, що з ростом покоління мобільної мережі, вимоги до нижньої межі затримки даних збільшуються. Також аналіз вимог до загальної мережевої затримки 5G показав, що вона повинна бути менше 1 мс.

На Рис. 6.2 представлено порівняння вимог до затримки на рівні управління та рівні користувача для сигнального та абонентського трафіків відповідно.

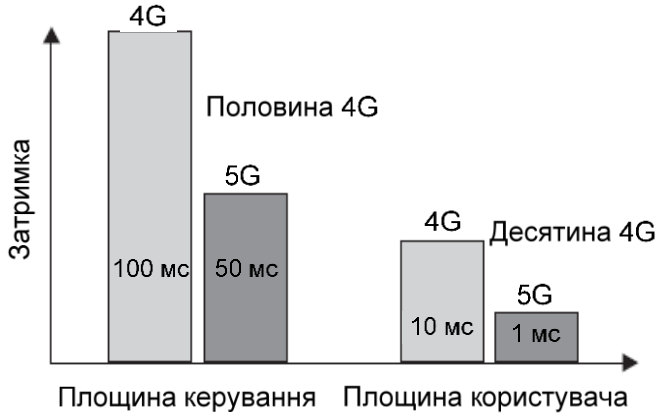


Рис. 6.2 Порівняння вимог до затримки на рівні управління та рівні користувача

На Рис. 6.2 видно, що вимоги до мережі 5G будуть в два рази жорсткішими для сигнального трафіку і в 10 разів жорсткішими для абонентського трафіку.

Іншим параметром є частка втрачених пакетів через помилки при отриманні пакетів даних (IP-packet Error Rate).

Значення цього параметра, що визначає вимоги максимального числа втрат IP пакетів для широкомовного відео за допомогою мобільних мереж 3G/4G/5G, наведені в Таблиця 6.2

Таблиця 6.2

Вимоги до кількості втрачених пакетів для широкомовного відео в 3G/4G/5G мережах

Терміни QoS	Запланована затримка пакетів (мс)		
	3G	4G	5G
Без гарантування якості	10^{-2}	10^{-3}	10^{-4}
З гарантованою якістю	10^{-2}	10^{-6}	10^{-7}

Для інших сервісів якість також буде визначатися часткою втрачених пакетів в мережах 3G/4G/5G. Умови обслуговування абонентських пристроїв будуть визначатися в обох випадках: з гарантованою якістю обслуговування і без гарантованої якості. Вимоги до коефіцієнта втрат пакетів для інших послуг наведені в Таблиця 6.3.

Таблиця 6.3

Вимоги до частки втрачених пакетів для інших сервісів в 3G/4G/5G мережах

Терміни QoS	Коефіцієнт втрат пакетів (Packet Error Loss Rate)			
	SDTV	HDTV	4k UHD	8k UHD
Покоління мобільних мереж	3G/4G	4G	4G	5G
Широкомовне відео з гарантованою якістю	10^{-6}	10^{-7}	10^{-8}	10^{-9}

Розвиток концепції NFV призведе до віртуалізації функцій управління якістю, які можуть бути розділені на 2 складові: управління (Cloud QoS management function - CQMF) і контролю (Cloud QoS control function - CQCF), які показані на Рис. 6.3.

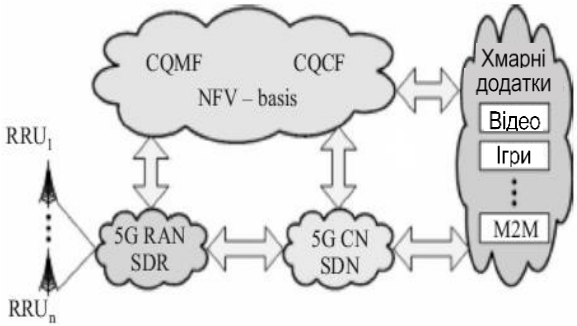


Рис. 6.3 Основні складові функцій управління якістю

Функція CQCF забезпечує контроль потоків трафіку в режимі реального часу на основі QoS, встановленого під час з'єднання. Основні механізми управління QoS включають управління профілями трафіку, планування і управління потоками даних.

Функція CQMF забезпечує підтримку QoS в мережі 5G відповідно до SLA, а також забезпечує моніторинг, технічне обслуговування, аналіз і масштабування QoS.

Реалізація алгоритмів пріоритетності трафіку в мережах 5G буде ґрунтуватися на процедурах класифікації трафіку з акцентом на пріоритети відео-трафіку і M2M трафіку. Класифікація процедур трафіку повинна бути проведена з урахуванням можливостей адаптації: характеристики трафіку будуть динамічно змінюватися з появою нових додатків, як в M2M області, так і в області відео.

6.2 Контроль якості у системі з віруалізованою системою керування

6.2.1 Забезпечення якості в мережах LTE без віртуалізації

Оркестрація та менеджмент (O&M) базової станції UMTS (Node B) Node B розділена на дві частини: O&M, пов'язана з фактичною реалізацією базової станції, позначена як конкретна реалізація (Implementation Specific) O&M, та O&M, що впливає на ресурси передачі трафіку в базовій станції, яка управляється контролером радіомережі, позначена як логічна O&M. Архітектура контролера мережі (RNC) з інтерфейсами O&M показана на Рис. 6.4.

Рис. 6.4 Архітектура контролера мережі з інтерфейсами O&M

На функції конкретної реалізації O&M значною мірою впливає реалізація базової станції (NodeB), як з точки зору її апаратних компонентів, так і з точки зору управління компонентами програмного забезпечення. Тому функції є залежними від реалізації та виконуватися між базовою станцією та системою управління.

Як варіант транспортного рішення для конкретної реалізації O&M - це маршрут від базової станції до системи управління через контролер мережі RNC. У цьому випадку інтерфейс конкретної реалізації O&M для та інтерфейс Iub мають одного і того ж фізичного каналу, а TS 25.442 [4] визначає функцію маршрутизації та транспортного каналу для цього сценарію. Розгортання маршрутизації через RNC в UTRAN є необов'язковим. Там, де потрібна сигналізація між спільно розміщеним обладнанням та його системою управління, вона може бути перенесена на той самий канал, що і конкретна реалізація O&M.

Логічна O&M - сигналізація, пов'язана з контролем логічних ресурсів (каналів, сот, ...), що належать контролеру мережі RNC, але фізично реалізовані у базовій станції. Контролер радіомережі контролює ці логічні ресурси. Ряд процедур O&M, фізично реалізованих у Node B, впливають на логічні ресурси, і тому вимагають обміну інформацією між RNC та Node B. Усі повідомлення, необхідні для підтримки цього інформаційного обміну, класифікуються як логічні O&M, які є невід'ємною частиною сигнального протоколу NBAP (Node B Application Part).

6.2.2 Вимоги до NFV і SDN інтеграції в мобільних мережах

Основні вимоги до гетерогенних мереж поряд зі скороченням витрат операторів полягають в задоволенні необхідного значення показника доступності, який в свою чергу включає показники:

- відмовостійкості;
- підвищеної продуктивності;
- низькою затримки.

Крім цього потрібна наявність вбудованого моніторингу, забезпечення чіткого шляху міграції, сумісності з успадкованими системами, забезпечення актуального переліку послуг, що надаються.

Вимоги до рівня доступності сучасних мобільних мереж розглянуті в роботі [3] і автори відзначають, що значення рівня доступності має бути близько 99,999%, цей показник повинен бути збережений або покращений.

Інтеграція SDN та віртуалізація мережевих функцій зводить до мінімуму зміни в мережевих елементах, забезпечуючи тим самим плавний перехід на основі потреб оператора. Безпека має першочергове значення для мобільних мереж і її слід розглядати для всіх рівнів, мережевих функцій, а також фізичних і віртуальних елементів. Починаючи від контролера SDN, який має доступ до всієї мережевої архітектури, закінчуючи вузлом, який виконує функції мережі, система повинна гарантувати критичний рівень безпеки і високої доступності.

Ще однією важливою вимогою є моніторинг мережі. Моніторинг мережі полегшує перевірку і підтвердження Service Level Agreements (SLA), продуктивності (якість обслуговування – QoS), пошуку та усунення несправностей, а також оцінку та оптимізацію використання ресурсів. З одного боку, віртуалізація мережевих функцій

встановлює нові вимоги до моніторингу мобільного мережі, але з іншого боку також надає засоби для реалізації передових рішень моніторингу мережі. NFV/SDN дозволяє інтегрувати хмарні інфраструктури, що забезпечують більш високу ступінь свободи щодо розміщення точок вимірювання та гнучкого управління потоками трафіку. Передове і ефективне рішення моніторингу QoS має включати як розподілену (на основі SDN/NFV) систему вимірювання QoS, так і централізовану систему оцінки.

Налаштування сервісів і оптимізація – ще одні вимоги для забезпечення доступності ресурсів. Це може бути зроблено за допомогою оркестратора. Ця вимога може бути розгорнута в SDN мережах, використовуючи додатки управління, які мають повне уявлення про конфігурацію мережі. Разом з інформацією про статус від систем моніторингу мережі та збору даних, це дозволяє оркестратору мобільної мережі оптимізувати обслуговування (наприклад, затримку) і/або використовувати менше ресурсів, ніж традиційні мережі. Оркестратор може контролювати кілька мережевих елементів через додатки управління. Це дозволяє впроваджувати нові послуги вносячи зміни в оркестратор, тоді як в традиційних мережах для підтримки нових сервісів необхідно оновлювати все обладнання. Мереж на базі SDN повинні співіснувати з наявною архітектурою мережі. Для того, щоб задіяти потенціал SDN, потрібна взаємодія цих двох рішень, наприклад, шляхом введення рівня абстракції і автоматизації в застарілу мережеву частину.

6.3 Показники якості послуг передачі даних

В системах мобільного зв’язку третього покоління функції контролю параметрів доступності сервісів виконувала підсистема SGSN, в ядрі EPC ці функції виконує підсистема PCRF.

На Рис. 6.5 зображена архітектура мережі передачі даних 2G/3G. Основним мережевим вузлом в цій архітектурі є вузол підтримки послуг GPRS (Serving GPRS Support Node, SGSN), який відстежує розташування абонентських терміналів передачі пакетних даних, забезпечує захисні функції і контроль доступу. SGSN з'єднується з контролером базових станцій BSC, використовуючи Frame Relay - мережу комутації фреймів (пакетів другого, канального рівня). Вузол шлюзовий підтримки GPRS (Gateway GPRS Support Node, GGSN) взаємодіє з зовнішніми мережами пакетної передачі даних (Packet Data Networks, PDNs), забезпечуючи передачу даних до/від мобільних терміналів. GGSN пов'язаний із зовнішніми мережами з пакетним перемиканням і з'єднується з вузлами SGSN через мережі, які використовують протокол IP. Керуючий блок пакетної комутації (Packet Controller Unit, PCU) забезпечує послуги пакетної радіопередачі на область охоплення BSC. Для з'єднання вузлів SGSN і GGSN з іншими елементами мережі, що забезпечують глобальну пакетну передачу даних, може додаватися кілька нових інтерфейсів (з позначенням G *, де * – символ, що визначає конкретний інтерфейс).

Розглянемо основні причини збою в роботі системи передачі даних. Однією з найбільш поширених причин є помилка, яка відправляється з боку GGSN на Gn інтерфейс. Ця помилка широко поширена в GTP повідомленнях інформаційного елемента.

PDP Context – набір даних про абонента, що здійснив процедуру GPRS Attach (процедуру аутентифікації і надання доступу в інтернет), який зберігається як на стороні SGSN, так і на стороні терміналу абонента, зокрема в цей набір даних входить профіль, що забезпечує певний рівень якості обслуговування абонента – QoS, призначена абоненту адреса в мережі, деякі дані про тарифікацію абонента.

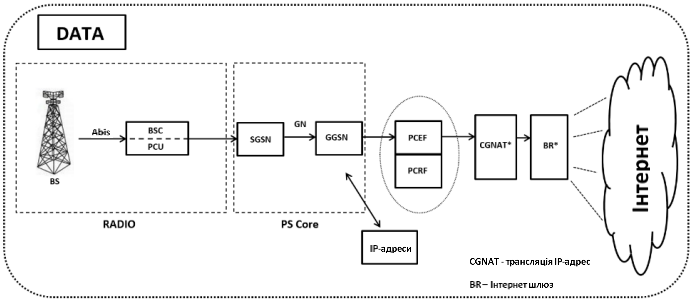


Рис. 6.5 Архітектура мережі передачі даних 2G/3G

Розглянемо основні показники якості сервісу передачі даних

1. Доступність

• Відсоток успішно виконаних процедур активації PDP context, ініційованих MS 2G/3G (PDP Context Activation Success Rate)

$$PDPctx\_act\_SR = \frac{PDPctx\_act\_acc}{PDPctx\_act\_req} * 100, \text{ де}$$

PDPctx_act_acc - число успішно виконаних процедур активації PDP-контексту, ініційованих MS (таймаут = 150с, після чого спроба вважається неуспішною).

PDPctx_act_req - число запитів на виконання процедури активації PDP-контексту, ініційованих MS.

• Затримка часу встановлення TCP-сесії для сервісу Web Browsing

$$WebBrow\,sin\,g \text{ Latency} =$$
$$\frac{1}{N}\sum(\text{WBST-WBRT})_i, \text{ де}$$

WBST(WebBrowsingSuccessTime) – час встановлення TCP-сесії для сервісу Web Browsing;

WBRT(WebBrowsingRequestTime) – час запиту встановлення TCP-сесії для сервісу Web Browsingі - і-ий запит встановлення TCP-сесії

N - сумарна кількість запитів за інтервал спостереження.

Відсоток блокувань в ЧНН в режимі передачі даних через перевантаження на 2G/3G (Відношення між кількістю відмов у виділенні ресурсів для передачі даних через перевантаження і загальною кількістю запитів на виділення ресурсів в ЧНН 2G/3G (Connection Block Rate).

$$ConnectionBlockRate = (PS\_BlockRate2G * Traffic2G + PS\_BlockRate3G * Traffic3G)/(Traffic2G + Traffic3G)$$

PS_BlockRate2G – відсоток блокувань в ЧНН в режимі передачі даних через перевантаженяь для мережі 2G

PS_BlockRate3G – відсоток блокувань в ЧНН в режимі передачі даних через перевантаженя для мережі 3G

$Traffic2G$ – пакетний трафік в мережі 2G в ЧНН, MB

$Traffic3G$ – пакетний трафік в мережі 3G в ЧНН, MB

• Відсоток успішно виконаних спроб реєстрації мобільної станції в мережі пакетної передачі даних 2G/3G (PS Attach SR).

$$PS\_Attach\_SR2G/3G = (PS\_Attach\_SR2G * Traffic2G + PS\_Attach\_SR3G * Traffic3G)/(Traffic2G + Traffic3G)$$

PS_Attach_SR2G – відсоток успішно виконаних спроб реєстрації мобільної станції в мережі пакетної передачі даних 2G;

PS_Attach_SR3G – відсоток успішно виконаних спроб реєстрації мобільної станції в мережі пакетної передачі даних 3G;

$Traffic2G$ – середньодобовий пакетний трафік в мережі 2G, MB;

$Traffic3G$ – середньодобовий пакетний трафік в мережі 3G, MB;

• Відсоток успішно виконаних процедур активації PDP context, ініційованих MS 2G/3G (PDP Context Activation Success Rate)

$$PDPctx\_act\_SR2G/3G = (PDPctx\_act\_SR2G * Traffic2G + PDPctx\_act\_SR3G * Traffic3G)/(Traffic2G + Traffic3G)$$

$PDPctx\_act\_SR2G$ – відсоток успішно виконаних процедур активації PDP context, ініційованих MS в 2G

$PDPctx\_act\_SR3G$ – відсоток успішно виконаних процедур активації PDP context, ініційованих MS в 3G

$Traffic2G$ - середньодобовий пакетний трафік в мережі 2G, MB

$Traffic3G$ - середньодобовий пакетний трафік в мережі 3G, MB

• Стабільність роботи центральних систем пропуску data-трафіку

$$data\_traf\_stability = 1 - \frac{t_{traf\_unstable}}{t}$$

ttraf_unstable – час нестабільної роботи систем пропуску трафіку вважається період, протягом якого рівень трафіку на PS-Core був нижче більш ніж на 20% рівня відповідного періоду минулого тижня. Показник стабільності роботи центральних систем пропуску трафіку є відношенням часу, протягом якого системи працювали стабільно, до загального часу періоду t.

2. Цілісність

• Частка успішних тестів WebBrowsing з часом закачування сторінки Kepler розміром 800 Кбайт не більше 130 секунд,

$$Activities\_WBtime < 130sec = \frac{Activities\_WBtime<130sec}{\text{TotalActivities}} * 100\%, \text{ де}$$

$Activities\_WBtime < 130sec$ - кількість абонентських сесій з WebBrowsing Time <130sec, TotalActivities - загальна кількість абонентських сесій за період вимірювань.

• Середня швидкість передачі даних на одну абонентську активність для сервісу Web Browsing

$$WebBrowsing\ AverageSpeed = \frac{1}{N}\sum \text{WebBrowsingSpeed}_i, \text{, де}$$

i - i-а абонентська активність WebBrowsing-а

N - сумарна кількість запитів сервісу WebBrowsing.

• Частка абонентських сесій, які отримують дані при використанні VS із середньою швидкістю понад 400 кбіт/с

$$Activities\_VS > 400\text{kbps} = \frac{Activities\_VS > 400\text{kbps}}{\text{TotalActivities}} * 100\%, \text{ де}$$

$Activities_VS > 400$kbps- кількість абонентських сесій з VideoStreaming Speed> 400kbps,

TotalActivities - загальна кількість абонентських сесій за період вимірювань.

6.3.1 Аналіз методів забезпечення параметрів QoS в сервісно-орієнтованій архітектурі LTE

Концепція системи QoS для мереж UMTS мобільного зв'язку 3-го покоління визначена в специфікації TS 23.107, і використовується також для мереж LTE 4-го покоління [118, 119, 120].

На Рис. 6.6 розглянуто архітектуру системи якості обслуговування і передавання послуг у рамках такої системи, для випадку, коли зв'язок здійснюється між кінцевим обладнанням (КО), підключеним до призначеного для користувача терміналу мобільної мережі, і термінальним обладнанням, розташованим в зовнішній пакетній мережі.

Аналогічно поняттю наскрізного каналу вводиться поняття *наскрізної послуги* (end–to–end service) як послідовність дій між двома кінцевими користувачами і, відповідно, частин послуг – по їх відношенню до визначених мережевим складовим: в локальному каналі «КО – призначений для користувача термінал» (Terminal Equipment/Mobile Terminal local Bearer Service), в каналі мережі LTE (LTE Bearer Service), в зовнішньому каналі (External Bearer Service). Таким чином, виникає багаторівнева взаємодія при передачі послуги в різних мережевих вузлах і на різних рівнях.

Передача послуги з мережі LTE розглядається, відповідно до мережевої архітектури, окремо в мережі радіодоступу (Radio Access Bearer Service), де забезпечується конфіденційне передавання призначених для користувача даних або із заздалегідь вибраним або встановленим за замовчуванням рівнем якості обслуговування, і у базовій пакетній мережі (Core Network Bearer Service), що також може підтримувати різну якість обслуговування.

Послугу в мережі радіодоступу реалізують двома частинами: в радіоканалі (Radio Bearer Service) і в механізмі радіодоступу (Access Bearer Service). Реалізація послуги в радіоканалі містить усі аспекти, що стосуються передавання даних по радіоінтерфейсу, включаючи сегментацію і повторне збирання призначених для користувача пакетів. Крім того, на фізичному рівні (Physical Radio Bearer Service) здійснюється управління підпотоком призначених для користувача даних. Механізм радіодоступу забезпечує на фізичному рівні (Physical Bearer Service) передавання даних між мережею радіодоступу і

базовою мережею.

Рис. 6.6 Архітектура системи якості обслуговування [121]

Нарешті, проходження послуги в «магістральному» каналі (Backbone Network Bearer Service) розглядається у функціональній сукупності рівнів 1 і 2 і призначених вимог якості обслуговування.

Основні функції мережі LTE, що належать до керування якістю обслуговування в площині користувача

Функція відображення (MF, Mapping Function) забезпечує маркування кожного призначеного для передавання пакету даних відповідними параметрами QoS.

Функція класифікації (CF, Classification Function) призначена для виставляння пакетам параметрів QoS, призначених для певного АТ, у тому випадку, якщо для цього АТ в мережі встановлено декілька каналів передавання послуг.

Функція управління ресурсами (RMF, Resource Manager Function) розподіляє доступні ресурси між послугами відповідно до параметрів QoS.

Функція узгодження (очищення) трафіку (TCF, Traffic Conditioner Function) забезпечує узгодження між потоком призначених для користувача даних і встановленим рівнем якості обслуговування. Ті пакети даних, які не відповідають виставленим

параметрам QoS, будуть відкинуті або помічені як невідповідні для наступного відкидання після накопичення.

На **Ошибка! Источник ссылки не найден.** показана взаємодія функцій керування якістю обслуговування в призначеній для користувача площині [122].

Функція класифікації, реалізована в абонентському терміналі АТ і сигнальному шлюзі СШ, призначає пакети даних, отримані із зовнішнього (чи локального) каналу в послугу мережі LTE з відповідними параметрами QoS. Функція узгодження трафіку, при необхідності, забезпечує узгодження призначеного для користувача потоку у висхідному (у АТ) і низхідному (у СШ) напрямах зі встановленими параметрами QoS. Далі функція відображення забезпечує кожен пакет даних спеціальним QoS-індикатором, відправляючи того в мережу, що вимагає виділення відповідних ресурсів – за це відповідальна функція управління ресурсами, реалізована в кожному мережевому вузлі.

У площині управління зосереджені функції, необхідні для реалізації механізмів управління і контролю.

Основні функції мережі LTE, що належать до керування якістю обслуговування в площині керування

Функція управління послугами (SMF, Service Manager Function) є координуючою функцією при установці, модифікуванні і керуванні послугами, а також що координує функції керування якістю обслуговування в призначеній для користувача площині.

Функція (TF, Translation Function) *трансляції* перетворює внутрішні примітиви послуг мережі LTE в модулі різних протоколів взаємодіючих зовнішніх мереж, включаючи перетворення атрибутів послуг мережі LTE в параметри QoS протоколів зовнішніх мереж.

Функція управління можливостями (A/CCF, Admission / Capability Control Function) забезпечує інформацією про усі можливі ресурси мережевих вузлів, визначаючи при кожному запиті (чи модифікуванні) послуги, чи можуть мережеві вузли забезпечити необхідні ресурси. Ця функція також контролює можливість надання самої послуги, тобто чи реалізована в мережі запитана послуга.

Функція керування підпискою (SCF, Subscription Control Function) забезпечує контроль доступності абонентам певних послуг з необхідними параметрами QoS.

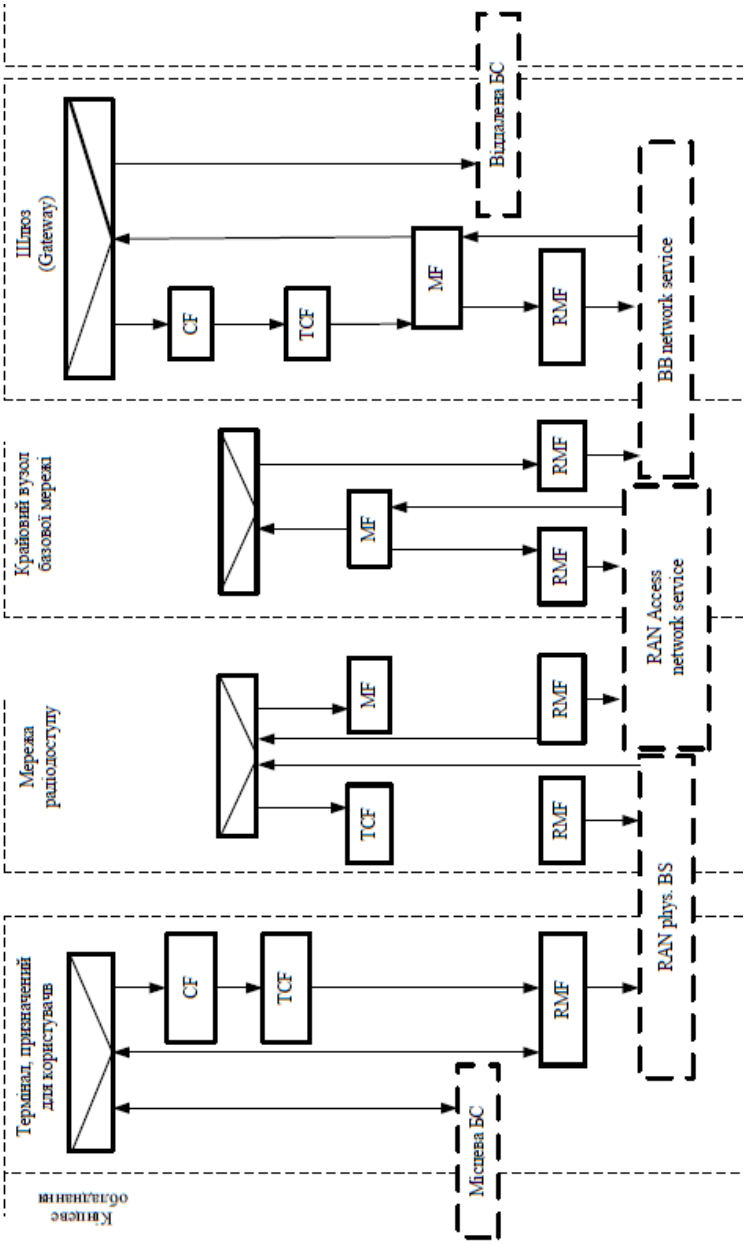


Рис. 6.7 Функції керування якістю обслуговування в призначеній для користувача площині

6.3.2 Реалізація концепції QoS у мережах LTE

Відповідно до рекомендації 3GPP TS 23.402 V15.1.0 (2017-09), модель QoS, яка застосовується в поєднанні з опорними точками на основі протоколу PMIP, не використовує ідентифікатори каналів у пакетах площини користувача. Вона базується на фільтрах пакетів та пов'язаних з ними параметрах QoS (QCI,

ARP, MBR, GBR), що надаються системі доступу через позамаршрутну (off-path) сигналізацію.

PCRF сигналізує ті ж фільтри пакетів та пов'язані з ними параметри QoS через Gxa, Gxb та Gxc, що і через Gx; інакше кажучи, деталізація інформації QoS, яка передається через Gxa, Gxb та Gxc, така ж, як і через Gx.

Канал EPS з S5/S8 на основі протоколу PMIP і протоколу доступу E-UTRAN показано на Рис. 6.7.

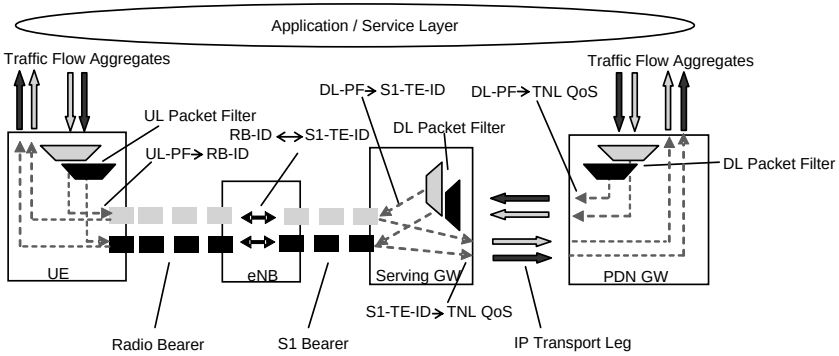


Рис. 6.7 Два канали Unicast EPS (S5/S8 на основі PMIP та протокол доступу E-UTRAN)

Для S5/S8 на основі протоколу PMIP та протоколу доступу E-UTRAN, канал EPS містить конкатенацію одного радіоканалу та одного каналу S1. Служба підключення PDN між обладнанням користувача (UE) і зовнішньою мережею пакетної передачі даних підтримується шляхом об'єднання каналу EPS та IP-з'єднання між шлюзом обслуговування і шлюзом PDN. Контроль QoS між Serving GW і PDN GW забезпечується на рівні транспортної мережі (TNL).

Канал EPS реалізується наступними елементами:

- В обладнанні користувача UFT TFT відображає агрегат потоку трафіка до каналу EPS у напрямку висхідної лінії зв'язку.
- У шлюзі обслуговування (Serving GW) DL TFT відображає агрегат потоку трафіку до каналу EPS у напрямку низхідної лінії зв'язку.
- радіоканал передає пакети каналу EPS між UE та eNodeB. Існує одноразове відображення між каналом EPS та радіоканалом.
- Канал S1 переносить пакети каналу EPS між eNodeB і шлюзом обслуговування. Існує одноразове відображення між каналом EPS та каналом S1.
- Для кожного обладнання користувача в туннелі PDN транспортуються пакети каналу EPS між шлюзом обслуговування та шлюзом PDN. Між каналом EPS та цим тунелем реалізується відображення «багато до одного».

- UE зберігає відображення між фільтром пакетної передачі висхідної лінії зв'язку та радіоканалом для створення зв'язків між агрегатом потоку трафіку та радіоканалом у висхідній лінії зв'язку.
- eNodeB зберігає відображення «один до одного» між радіоканалом та каналом S1, щоб створити зв'язок між радіоканалом та каналом S1 як у напрямку висхідної лінії зв'язку, так і в напрямку низхідної лінії зв'язку.
- Шлюз обслуговування зберігає відображення «один до одного» між фільтром пакетної передачі по низхідній лінії зв'язку та каналом S1, щоб створити відображення між агрегатом потоку трафіку та каналом S1 у низхідній лінії зв'язку.
- Шлюз доступу до інших мереж (PDN SW) забезпечує APN-AMBR для всіх SDF тих самих APN, які пов'язані з QCI без GBR.

6.3.3 Процедури виділеного каналу для протоколу доступу E-UTRAN з S5/S8 на основі PMIP

Процедура, зображена на рисунку

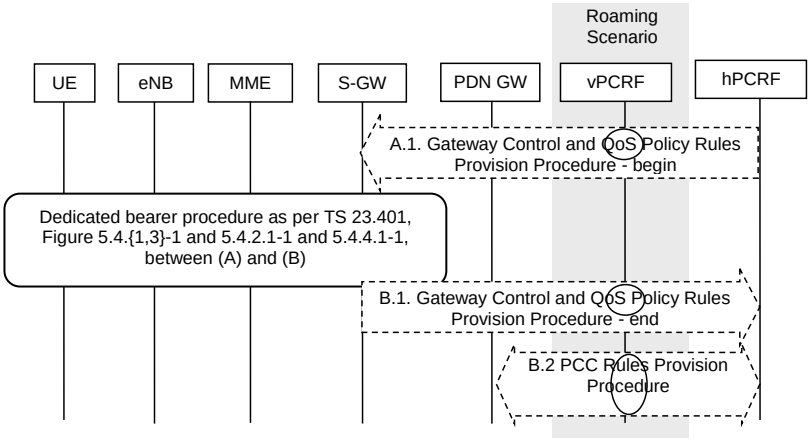


Рис. 6.8,

застосовується до всіх виділених операцій розподілу ресурсів для E-UTRAN, які активуються за допомогою PCRF, за винятком процедури дезактивації виділених каналів, ініційованої MME. Процедури, ініційовані по S-GW в E-UTRAN відрізняються для кожного випадку.

Процедура, описана на

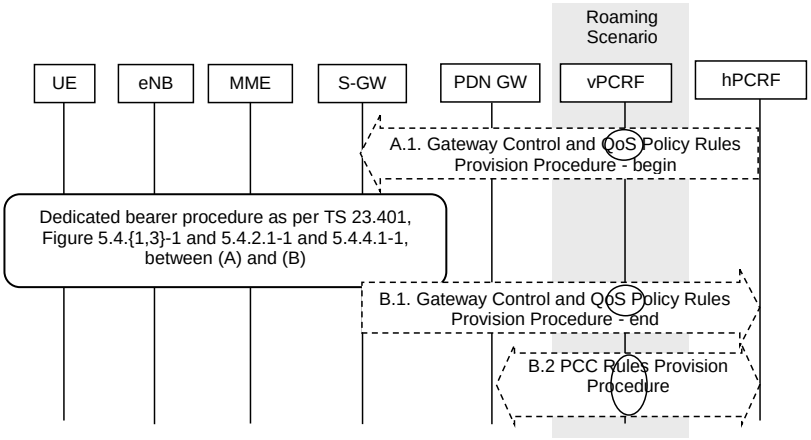


Рис. 6.8, показує лише етапи, пов'язані з використанням S5/S8 на основі PMIP, які відрізняються від варіанту процедури GTP, наведеної в TS 23.401 [4].

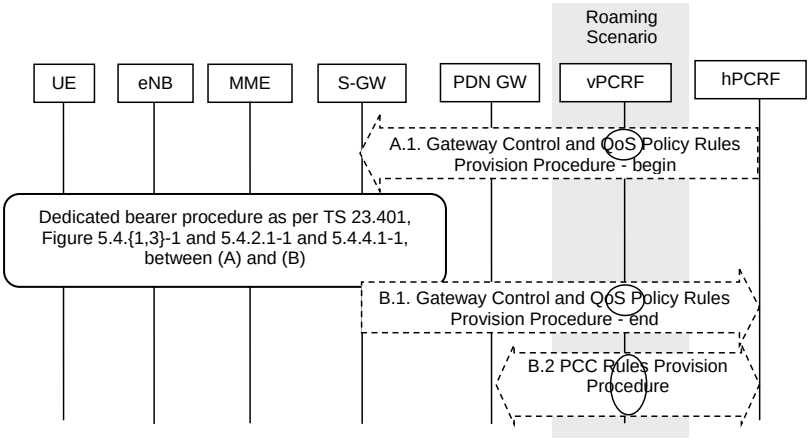


Рис. 6.8

Процедура розподілу ресурсів, UE в активному режим [4]

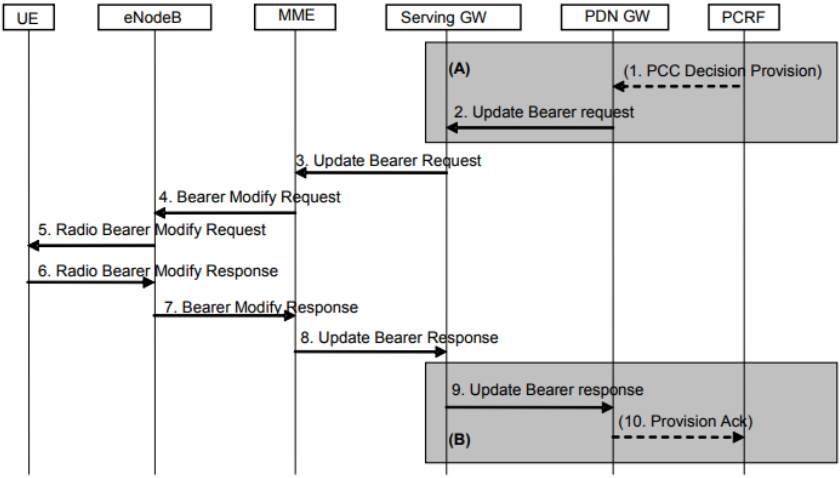


Рис. 6.9 Процедура модифікації носіїв з оновленим QoS носієм, UE в активному режимі

Ця процедура застосовується до випадків не роумінгу, роумінгу та місцевої обробки запитів (Local Breakout). Для випадків роумінгу та Local Breakout vPCRF пересилає повідомлення між шлюзом обслуговування і hPCRF. У разі Local Breakout vPCRF пересилає повідомлення між PDN GW та hPCRF.

Якщо динамічне положення політики не розгорнуто, кроки, наведені на малюнку, не виконуються. Замість цього може бути застосована налаштована статична політика.

A.1) PCRF ініціює процедуру постачання правил Gateway Control і QoS, зазначену в TS 23.203 [123], відправляючи повідомлення з правилами QoS та інформацією тригера подій до S-GW.

Кроки між A.1 і B.1 описані в TS 23.401 [4].

Для S5/S8 на основі PMIP, перед етапами процедури (крок 3 TS 23.401 [4], п. 5.4.1), PCRF надсилає повідомлення про прийняття рішення щодо керування політикою та тарікацією PCC (Policy and Charging Control) (політика QoS) до S-GW, а не до P-GW, як зроблено для S5/S8 на основі GTP. S-GW використовує цю QoS-політику, щоб визначити, що потоки трафіку повинні бути об'єднані або вилучені з активного каналу. S-GW генерує TFT і оновлює якість обслуговування EPS-каналу, щоб відповідати агрегованому набору потоків трафіку. Функція зв'язування каналу S-GW може виконувати модифікацію, створення або видалення каналів у цій точці. Для модифікації S-GW надсилає MME повідомлення про оновлення запиту каналу (Update Bearer Request). Для створення виділеного каналу S-GW надсилає повідомлення «Створити канал» (Create Bearer), а для видалення, S-GW надсилає запит на видалення каналу (Delete Bearer Request).

B.1) Шлюз обслуговування вказує PCRF, чи може бути застосований запит на

порядок правил політики QoS (QoS Policy Rules Provision), таким чином доповнюючи процедуру GW Control та QoS Rules Provision, розпочату на кроці A.1. Шлюз обслуговування повідомляє PCRF про місцезнаходження користувача та/або про часовий пояс UE (UE Time Zone) у якості звіту про події (Event Report), якщо встановлено відповідний тригер події.

B.2) PCRF ініціює процедуру надання правил PCC, як зазначено в TS 23.203 [123]. PCRF надає оновлені правила PCC для PCEF для виконання шляхом процедури PCC Rules Provision, зазначеної в TS 23.203 [123].

Крок B.2 може виконуватись перед кроком A.1 або паралельно з кроками A.1 B.1, якщо для оновлення правил PCC в PCEF не вимагається підтвердження розподілу ресурсів. Детальніше описано в TS 23.203 [123].

6.4 Вплив віртуалізації мережевих функцій на процедури забезпечення якості обслуговування.

Незважаючи на успіхи, досягнуті в області розробки мереж мобільного зв'язку четвертого покоління, з'являються нові вимоги, викликані зростаючими потребами в комунікаціях, що в свою чергу вимагає розвитку нового покоління мобільних мереж (5G). Нові можливості використання, такі як потокове відео високої роздільної здатності, віддалений моніторинг, управління в реальному часі створюють вимоги, пов'язані з пропускною здатністю, затримкою передачі, надійністю і стійкістю мережі.

Очікується, що мережі забезпечать високу безпеку, мінімальну затримку (нижче 5 мс.), надійність передачі 99,999% і стовідсоткову доступність.

Розробка мережі 5G повинна бути спрямована на подолання вищевказаних обмежень, з метою надання ультра-надійних, безпечних сервісів з мінімальною затримкою великій кількості інтелектуальних об'єктів і систем, а також новим мобільних терміналів.

Порівняння технологій 3G/4G/5G представлено на Рис. 6.10. Технології IMT-2000 (3G) IMT-Advanced (4G) покращують пропускну здатність мережі, швидкість передачі даних користувачів, використання спектра і зменшують затримки. Впровадження технології IMT (5G) планується для «потужної і критично важливої машинної взаємодії», зменшення затримки, спектральної ефективності, швидкості, мобільності і надійності.

Мережева архітектура 5G складається з двох шарів: радіомережі та хмарної мережі. Різні типи базових станцій, що виконують мінімальний набір функцій, утворюють радіомережу. Хмарна мережа складається з площини користувача (User Plane Entity - UPE) і площині управління (Control Plane Entity - CPE), які виконують функції площини користувача і площини управління відповідно. Як показано на **Ошибка! Источник ссылки не найден.**, фізична реалізація хмари може бути адаптована для задоволення різних цільових показників. Наприклад, UPE і CPE можуть бути розташовані близько до базових станцій для зменшення затримки критичних послуг. Також може бути виконано підключення БС до невеликого прилеглого центру обробки даних (ЦОД-3), а не до центрального ЦОД-2. З іншого боку, БС може бути з'єднана з ЦОД-2, якщо затримка не критична. Така гнучкість дозволяє оператору розгорнути великі і малі центри обробки даних для підтримки конкретних потреб у послугах. Така архітектура спрощує мережу і забезпечує швидке і гнучке розгортання і управління.

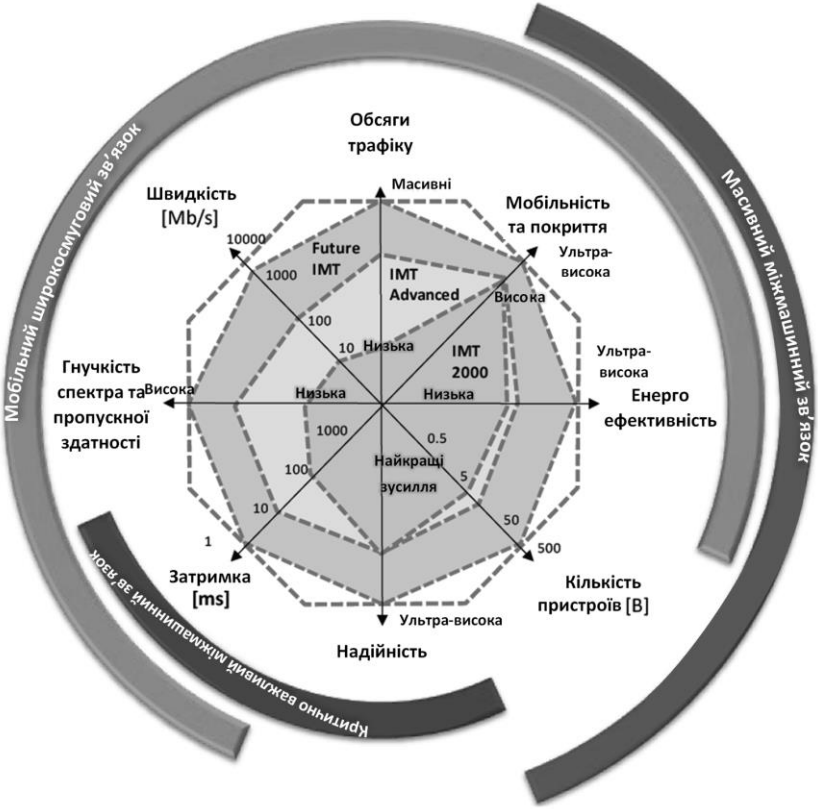


Рис. 6.10 Порівняння технологій 3G/4G/5G

Різні технології, такі як Network Functions Virtualization (NFV) і Software Defined Networking (SDN) призначені для створення і впровадження таких мереж. Проте, майбутні сервіси, такі як потокове відео та інші послуги мають різні вимоги, які підкреслюють необхідність динамічного масштабування функціональності мережі.

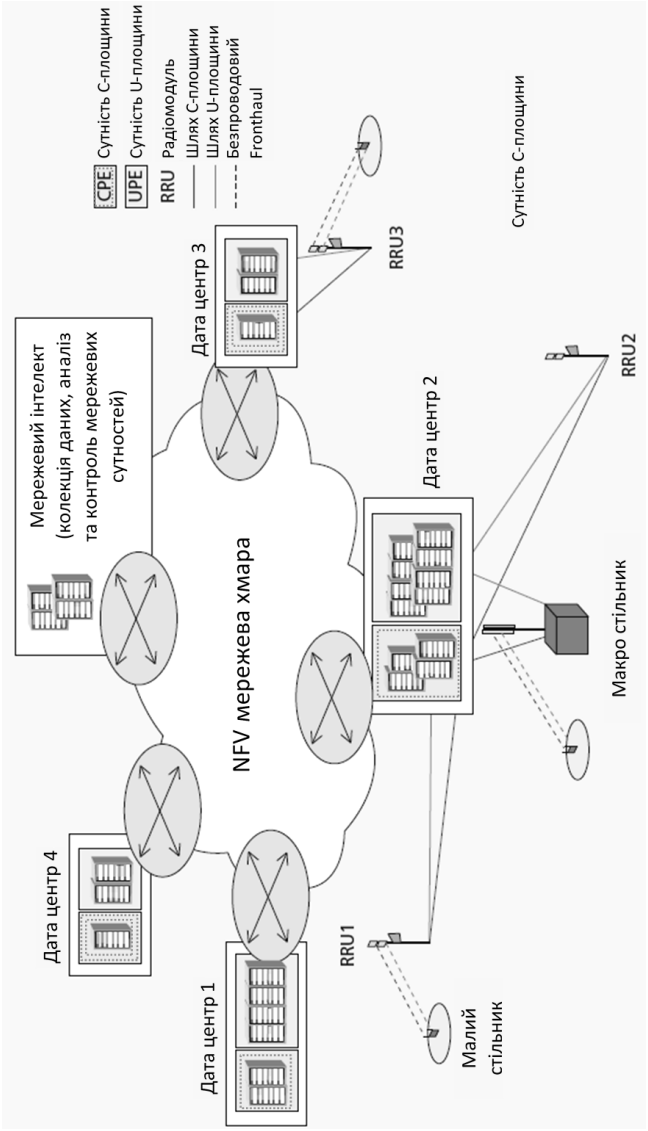


Рис. 6.12 Мережева архітектура 5G

Гетерогенні мережі відрізняються підвищеною пропускною здатністю, великим покриттям і надійністю, більш високою ефективністю використання радіочастотного спектру і низьким енергоспоживанням з боку як мережевої інфраструктури, так і термінального обладнання.

Мобільні оператори мережі стикаються з наростаючою проблемою зростання трафіку передачі даних, в зв'язку з поширеністю смартфонів і транслюванням аудіо і відео послуг. У новій парадигмі оператори повинні управляти навантаженням, задовольняючи зростаючі споживчі і корпоративні очікування продуктивності, забезпечуючи повсюдний широкосмуговий доступ, а також швидко впроваджувати нові послуги, щоб зберегти конкурентну перевагу. Існуючі мобільні мережі постійно стикаються з такими обмеженнями, як стаціонарне і дороге устаткування, складні протоколи управління і гетерогенні інтерфейси конфігурації. З метою вирішення поточних обмежень, необхідно вивчати і застосовувати принципи SDN в мобільних мережах, а саме SDMN (Software Define Mobile Networks). SDN розділяє рівні управління і передачі даних, використовуючи стандартні протоколи, що дозволяють віддаленим пристроям здійснювати управління і експлуатацію рівнів даних. Протокол синхронізації потрібен для комунікації обох рівнів, одним з таких протоколів є OpenFlow [2]. Переваги SDN в області хмарних обчислень є очевидними, однак застосування даної концепції в мобільних мережах вимагає подальшого вивчення.

6.5 Модель керування енергозбереженням телекомунікаційної системи, навантаженням та ресурсами обслуговування

В умовах невпинного розвитку інформатизації суспільства, постійного зростання одиниць електронних пристроїв та обладнання, що споживає електричну енергію виникла необхідність оптимізації систем керування процесом обслуговування в телекомунікаційних системах.

Зменшити споживання енергетичних ресурсів телекомунікаційною системою за рахунок керування навантаженням, процесом обслуговування в мережі радіо доступу, транспортній мережі та мережі ядра мобільного оператора, за рахунок впровадження комплексних моделей аналізу статистики, методів багатокритеріального оцінювання функціонування системи.

Телекомунікаційна мережа представлена у вигляді вузлів, вузлами телекомунікаційної мережі виступають різноманітні підсистеми, які мають різні технічні характеристики, а також виконують різноманітні функції такі як:

- Функції взаємодії з абонентськими терміналами (базові станції),
- Функції прийому та передачі інформаційних потоків у транспортній системі (комутаційне обладнання),
- Функції обслуговування службових інформаційних потоків (підсистеми ядра мобільного зв’язку).

Крім того вузли, що розглядаються також можна поділити на групи з подібними технічними характеристиками.

Метою поточних дослідних є зменшення енергетичних витрат на обслуговування телекомунікаційної мережі. На енергетичні витрати впливає низка факторів, які відрізняються залежно від функцій обладнання, виробника, та програмно-апаратного забезпечення телекомунікаційного вузла.

Однак доведено, що енергоспоживання залежить від навантаження на систему, тобто від кількості операцій які виконує пристрій. Тому організація потоків

навантаження на підсистеми телекомунікаційної мережі дозволить оптимізувати енергетичні витрати.

Суттєво зменшити витрати на обслуговування телекомунікаційної мережі доступу можна за рахунок зменшення потужності базових станцій. Зміна потужності базової станції призведе до низки пов'язаних з цим проблем. Однак на території міст, де покриття базовими станціями є дуже щільним, зменшення покриття однієї базової станції, що пов'язано зі зменшенням потужності суттєво не вплине на доступність послуг зв'язку, якщо параметри мережі базових станцій правильно розраховані.

Отже, постає задача: розрахувати параметрів для вузлів телекомунікаційної мережі таким чином, щоб загальне споживання електричної енергії було мінімальним, а якість надання послуг не змінилася.

Оскільки протягом доби навантаження на телекомунікаційну мережу змінюється, тому розрахунок та запровадження планів використання телекомунікаційного обладнання мобільного оператора дозволить зменшити витрати на енергоресурси, які забезпечують роботу телекомунікаційної мережі.

Постановка задачі:

N – кількість вузлів обслуговування, вузол обслуговування – група обладнання телекомунікаційної мережі, які обслуговують випадкове навантаження, що створює абонент, та може розглядатися як єдине ціле, та бути налаштоване відповідно вимогам моделі.

$N_д$ – кількість вузлів обслуговування, що утворюють мережу доступу.

$N_{тр}$ – кількість вузлів обслуговування, що утворюють транспортну мережу.

N_{core} – кількість вузлів обслуговування, що утворюють мережу ядра мобільного оператора.

$N1, N2, ..., N_M$ – підмножини вузлів обслуговування, на які розділяється множина N. Обладнання розташоване у вузлах, які входять в одну підмножину має одинакові характеристики.

M – кількість можливих груп обладнання.

наприклад, N1 – кількість базових станції, що підтримують технологію GSM,

N2 – кількість базових станцій, що підтримують технології 3G,

N3 – кількість радіорелейних станцій,

N4 – кількість SDH комутаторів

N5 – кількість комутаторів DWDM, та інші.

$$N = \sum_{m=1}^{M} Nm$$

Крім того, в рамках однієї підмножини *Nm (m=1,..M)* можна виділити обладнання від різних виробників, які можуть мати відмінні технічні характеристики та по різному споживати електричну енергію, відповідно потребують окремого налаштування. Тоді підмножина *Ni* розділяється на підмножини *Nms (m=1,..M, s=1,..S),* де S – максимальна кількість різних типів обладнання, які входять до однієї підмножини.

Слід зауважити, що N – велике число, наприклад, кількість базових станцій одного оператора перевищує 7000, і це число постійно зростає. Так само зростає кількість комутаційного обладнання мережі.

Залежно від типу вузла телекомунікаційної мережі будуть застосовані різні засоби енергозбереження. Для вузлів мережі доступу в результаті оптимізаційної роботи буде запропоновано план зміни потужності випромінювання, та можливе відключення. А для транспортної мережі буде запропонована схема границь навантаження, обслуговування якого є енергоефективним.

G – матриця зв'язаності вузлів. Якщо $i \in N_Д$ та $k \in N_Д$, тоді g_{ik} – показує наявність сусідства базових станцій.

J – кількість типів сервісів, які створюють навантаження на вузли телекомунікаційної мережі. В даній математичній моделі ми обмежуємося трьома типами трафіку – голосовий трафік, трафік даних та службовий трафік, в разі необхідності можна деталізувати.

t – змінна часу, дискретна величина, яка показує момент часу в який було зроблено спостереження.

$\lambda_{ij}(t)$ – математичне очікування інтенсивності навантаження на вузол *i* (*i=1,..N*) сервісом j-го типу (*j=1,..J*) в момент часу *t*. Для розрахунку математичного очікування аналізується статистика для однорідних періодів часу. Наприклад, математичне очікування навантаження заданого типу сервісу в будній день о 10 годині. При описі математичної моделі ми не зупиняємось на аналізі статистики, розрахунку середніх значень навантаження, та групуванні періодів часу де навантаження є однорідним. Далі розглядається модель типового робочого тижня.

$\lambda_i(t) = \sum_j \lambda_{ij}(t)$ – математичне очікування сумарного навантаження на вузол *i* (*i=1,..N*) в момент часу *t*.

Qj – множина граничних значень показників якості обслуговування для j-го (*j=1,..J*) типу сервісу. В результаті оптимізаційного процесу перерозподілу навантаження між вузлами телекомунікаційної мережі значення показників якості повинно зберегтися в межах заданих значень.

$q_{ij}(t)$ – значення показника якості *q* для *j*-го (*j=1,..J*) типу сервісу у узлі *i* (*i=1,..N*) в момент часу *t*

$\mu_{ij}(t)$ – інтенсивність обслуговування сервісу j-го типу (*j=1,..J*) у вузлі *i* (*i=1,..N*) в момент часу *t*. Міркування відносно часу аналогічні до параметру $\lambda_{ij}(t)$. Як було доведено, зміна параметру сумарної інтенсивності обслуговування $\mu_i(t) = \sum_j \mu_{ij}(t)$ призводить до зміни енергоспоживання *i*-го вузла.

$E_i(t)$ – енергоспоживання *i-го* (*i=1,..N*) вузла в момент часу *t*.

$f_i(\mu_i)$ – функції залежності енергоспоживання і-го (*i=1,..N*) від інтенсивності обслуговування сумарного навантаження на вузол. Структура функцій залежності розраховується експериментально для кожної з М груп обладнання.

Аналіз функцій $f_i(\mu_i(t))$ показав, що можна виділити діапазон значень $\mu_i(t)$ для яких енергоспоживання E_i буде оптимальнім.

$\lambda_{ij}'(t)$ – кількість запитів на обслуговування j-го типу (*j=1,..J*), які будуть надходити у вузол *i* (*i=1,..N*) в момент часу *t* після енергоефективного перерозподілу навантаження.

$\mu_{ij}'(t)$ – інтенсивність обслуговування сервісу j-го типу (*j=1,..J*) у вузлі *i* (*i=1,..N*) в момент часу *t* після енергоефективного перерозподілу навантаження.

Необхідно знайти розподіл навантаження $\lambda_{ij}'(t)$ та потужностей, що забезпечать інтенсивність обслуговування $\mu_{ij}'(t)$, при яких сумарне енергоспоживання телекомунікаційної мережі в момент часу *t*, $E(t) = \sum E_{i(t}(t)$, буде мінімальним, а показники якості обслуговування сервісів $q_{ij}(t)$ будуть збережені у межах заданих множиною *Qj*.

Розв'язок поставленої задачі є складною задачею, при розв'язанні якої пропонується користуватися наступними принципами:

1. Поетапно розв'язувати задачі оптимізації мережі доступу, транспортної мережі та для мережі ядра оператору зв'язку.

2. Врахувати джерела навантаження на вузли телекомунікаційної мережі.

3. Окремо розглядати вузли, для яких може бути застосовано спільний спосіб регулювання інтенсивності обслуговування.

4. Сформувати множини вузлів, між якими може бути перерозподілене навантаження.

Структура вхідного навантаження для мережі доступу. Навантаження на мережу доступу створюється мобільними пристроями, які зв'язуються з базовими станціями. Передача інформації від мобільного пристрою до базової станції можлива за умови наявності достатнього рівня сигналу від базової станції. Переключення між базовими станціями здійснюється за умови зниження рівня сигналу нижче допустимого значення.

Під зміною інтенсивності обслуговування для $i \in N_Д$ мається на увазі зміна потужності випромінювачів, які забезпечують downlink передачу даних до мобільного пристрою від базової станції. При зменшенні потужності базової станції зменшується радіус її дії, також знижується рівень сигналу базової станції, який бачить мобільний пристрій.

Окремою задачею для вузлів $i \in N_Д$ є *визначення нижньої границі потужності сигналу* відповідної базової станції, щоб покриття не змінилося.

Структура навантаження для транспортної мережі. Для вузлів $i \in N_Т$ вхідне навантаження має дві основні складові: навантаження, що надходить від вузлів мережі доступу $k \in N_Д$, яке може бути налаштовано при рішенні оптимізаційної задачі для мережі доступу, та транзитне навантаження, яке надходить від інших вузлів транспортної мережі. При використанні SDN технології є можливість рівномірно розподіляти інформаційні потоки по транспортній мережі, так щоб забезпечити енергоефективність роботи транспортних вузлів.

Під новим значенням інтенсивностей обслуговування $\mu_{ij}'(t)$ для вузлів $i \in N_Т$ розуміється розрахунок верхньої границі об'єму трафіку *j*-го типу сервісу (*j=1,..J*), що передається через вузол. При застосуванні технології програмно-керованих мереж можливе централізоване керування потоками у вузлах транспортної мережі. Відповідно об'єми трафіку, який проходить через транспортний вузол, можуть змінюватися залежно від часу доби, та будь-яких інших умов. Тобто енергозберігаюча схема транспортних потоків оператора зв'язку може бути застосована.

Енергоспоживання вузлів мережі ядра оператора зв'язку залежить від ефективності організації обчислювальних процесів. Обслуговування службових потоків у підсистемах ядра мобільного зв'язку, представляє собою послідовність стандартних операцій, виконання яких потребує обчислювальних ресурсів таких як оперативна

пам'ять, постійна пам'ять, процесорний час, та ресурс внутрішньої мережі. Обслуговування службових потоків здійснюється на декількох обслуговуючих одиницях (серверах), які розміщені територіально. На сьогоднішній день системи обслуговування службового трафіку тісно пов'язані з спеціальним обладнанням, де створюються службові потоки, та спеціальнім (вендорним) програмним забезпеченням, де вони обслуговуються. Однак поступовий перехід до віртуалізації окремих функцій та систем дозволить гнучке керування системою обслуговування службовими потоками. Навантаження на вузли $i \in N_{core}$ створюється при обслуговуванні службового трафіку. Проблеми обслуговування службового трафіку, оптимізація структури ядра оператора зв'язку буде розглянута нижче.

Далі запропоновано єдиний підхід до вирішення проблеми енергозбереження у вузлах телекомунікаційної мережі, який полягає в тому, щоб розподілити навантаження між доступними вузлами обслуговування, щоб забезпечити завантаження системи на заданому рівні, який забезпечує енергоефективне функціонування системи, крім того на заданому рівні зберігається показники якості обслуговування.

За основні прототипи математичної моделі, що зв'язує інтенсивність надходження запитів у вузол обслуговування телекомунікаційної мережі, інтенсивність обслуговування та показники якості функціонування обрано моделі К. Жернового, у роботах якого розглядаються системи масового обслуговування, де швидкість обслуговування пакету трафіку залежить від кількості заявок у черзі. Окремо розглядаються випадки з обмеженою та необмеженою чергою. Запропоновано розрахунок стаціонарного розподілу ймовірностей перебування системи у можливих станах, де під станом системи розуміється одночасне перебування в системі заданої кількості пакетів. В даній системі параметром керування виступає швидкість обслуговування пакету в залежності від кількості заявок у черзі. В результаті можна підібрати такий план швидкостей обслуговування, для яких результуючий розподіл буде задовольняти вимогам до показників якості обслуговування трафіку. Також є формули для розрахунку середньої зайнятості системи обслуговування.

Висновки

1. Запропонована методологія забезпечення показників якості обслуговування гібридного сервісу у гетерогенному телекомунікаційному середовищі базується на чотирьох основних підходах: формування навантаження на вузол обслуговування, вибір потужності вузла обслуговування, визначення порядку роботи вузлів обслуговування та поточний контроль ефективності роботи системи, що дозволило забезпечити ефективне функціонування гетерогенного середовища обслуговування гібридних телекомунікаційних сервісів, а саме зменшити відсоток запитів, що обслуговуються в системі більше допустимого часу, визначеного експертами, - на 2%, зменшити відсоток запитів, які було втрачено через перевищення допустимого часу обслуговування на 3%, зменшити відсоток часу коли ресурси завантажені: менше за допустиме значення - на 8% та більше за допустиме значення на 10%.

2. Запропоновано модель керування інфраструктурою NFV, яка враховує особливості розташування дата центрів обслуговування, а також топологічної структури потоків гібридних телекомунікаційних сервісів, які поступають на обслуговування у дата центри, дозволяє забезпечити гнучке керування інформаційно-телекомунікаційною системою організованою із застосуванням хмарних обчислень.

Контрольні запитання

1) Які механізми контролю якості обслуговування виділяють?
2) Які вимоги до якості існують в 3G/4G/5G мережах?
3) Назвіть основні складові функцій управління якістю.
4) Поясніть систему O&M.
5) Які основні вимоги до гетерогенних мобільних мереж?
6) Які основні показники якості сервісу передачі даних?
7) Які основні функції мережі LTE, що належать до керування якістю обслуговування в площині користувача?
8) Які основні функції мережі LTE, що належать до керування якістю обслуговування в площині керування?
9) Як реалізована концепція QoS у мережаж LTE?
10) Як реалізована процедура розподілу ресурсів у мережаж LTE?
11) Як виглядає мережева архітектура 5G?
12) Які різноманітні функції виконують вузли телекомунікаційної мережі?
13) Як зменшити енергетичні витрати у мережі?

РОЗДІЛ 7
ІМІТАЦІЙНІ МОДЕЛІ СИСТЕМ ОБСЛУГОВУВАННЯ

7.1 Дослідження ефективності методу оптимального вибору обчислювальних ресурсів для білінгових систем

Лояльність абонентів напряму залежить від рівня якості наданих послуг; пропозиції надання нових сервісів; швидкості роботи сервісів. Якість повинна, якщо не стабільно покращуватись, то утримуватись на достатньо високому рівні.

У сучасному світі об'єми інформації, що передається, невпинно збільшуються, що, у свою чергу, потребує постійного розширення пропускних можливостей мобільної мережі. Як наслідок, навантаження на внутрішні підсистеми обробки трафіку збільшуються, при цьому критично важливо уникнути перенавантажень на компонентах мережі, втрати даних і черг на обробку замовлень.

Дохід оператора мобільного зв'язку напряму залежить від організації процесу контролю фінансового стану рахунку (балансу) абонента - точності списання коштів абонента у режимі реального часу. Це дозволяє уникнути дебіторської заборгованості і підвищити комфорт користування послугами оператора.

Описані вище фактори визначають роль білінгової системи, як одного з головних компонентів в організації всього бізнес-процесу.

Якість послуг і дохід операторів мобільного зв'язку суттєво залежіть від того, наскільки правильно розрахована продуктивність та ефективність білінгових систем, що обробляють потоки вхідних заявок на обслуговування викликів тарифікації. Для білінгових підсистем, які надають послуги по списанню коштів абонента у режимі реального часу, важливою характеристикою роботи є обробка вхідних заявок в найкоротші терміни, без створення черги.

Щоб на вхідних інтерфейсах білінгової підсистеми обслуговування викликів тарифікації не виникали черги оброблюваних даних, необхідно контролювати об'єм вхідного потоку, підтримувати оптимальне навантаження і своєчасно планувати розширення білінгових підсистем, що обробляють такий потік.

Вибраний метод дозволяє провести розрахунок рекомендованого значення інтенсивності вхідного потоку на існуючій білінговій підсистемі, порівняти його з доступними метриками продуктивності і спланувати рекомендоване розширення білінгових модулей при прогнозованому збільшенні вхідного потоку.

7.1.1 Метод формування вхідного потоку навантаження для ефективного використання ресурсів обслуговування

Основна ідея методу полягає в тому, щоб виходячи з ергодичного розподілу для можливих станів системи сформувати вимоги до середнього значення вхідного навантаження, що дозволить максимально ефективно використовувати наявні фізичні ресурси обслуговування вхідного потоку заявок.

Процес обслуговування моделюється як n-канальний обслуговуючий пристрій, час обслуговування заявки у каналі є випадковою величиною розподіленою за законом Пуассона.

Вхідні дані.

n – кількість каналів для одночасного обслуговування заявок.

μ – інтенсивність обслуговування заявки,

G – кількість ресурсів залучених для обслуговування заявок,

v^g – об'єм g-го ресурсу необхідний для обслуговування у блоці однієї заявки, $g = \overline{1,G}$)

V^g – доступний об'єм ресурсу g-го ресурсу який спільно використовується заявками.

s – допустима кількість запитів у черзі на обслуговування.

R – відсоток заявок, які обслуговуються у системі не більше допустимого часу затримки, визначається експертами.

l – кількість запитів у черзі, до досягнення якої блокується надходження запитів до системи, відповідно до алгоритмів раннього попередження перевантажень.

Вихідні дані.

λ - рекомендоване значення для інтенсивності вхідного потоку, який буде направлено на обслуговування у s-канальний обслуговуючий пристрій.

Застосування запропонованого методу складається з двох етапів.

Етап 1. Для багатоканальної системи обслуговування, відповідно до моделей К. Жернового, необхідно знайти ергодичний розподіл кількості заявок у системі за формулами:

$$p_0 = \frac{1-\beta^{s-l}}{A_n(\alpha,\beta)}, \qquad \beta \neq 1, \qquad \alpha = \lambda/\mu, \qquad \beta = \lambda/n$$

$$A_n(\alpha,\beta) = (1-\beta^{s-l})\sum_{k=0}^{n}\frac{\alpha^k}{k!} + \frac{\alpha^n}{n!}\left(\frac{\beta-\beta^{s-l+1}}{1-\beta} - (s-l)\beta^{s+1}\right)$$

$$p_k = \frac{\alpha^k}{k!}p_0 \qquad (k = \overline{1,n})$$

$$p_{n+k} = \frac{\alpha^n}{n!}\beta^k p_0 \quad (k = \overline{1,l})$$

$$p_{n+k} = \frac{\alpha^n}{n!}\frac{\beta^k-\beta^s}{1-\beta^{s-l}}p_0 \; (k = \overline{l+1,s-1})$$

$$p_{n+s} = \frac{\alpha^n}{n!}\frac{(1-\beta)\beta^s}{A_n(\alpha,\beta)} \quad (k = \overline{n+l+1,n+s-1})$$

Якщо $\beta = 1, \quad \alpha = \lambda/\mu, \;\; \beta = \lambda/n,$ тоді

$p_k = \frac{n^k}{k!}p_0 \;\; (k = \overline{1,n}); \;\; p_{n+k} = \frac{n^k}{k!}p_0 \;\; (k = \overline{1,l});$

$p_{n+s} = \frac{n^n}{n!}p_0 \;\; (s-l-1)p_{n+s} \Longrightarrow$

$$\Longrightarrow p_{n+s} = \frac{n^n}{n!\,(s-l)}p_0 \;\; \left(k = \overline{n+l+1,n+s-1}\right);$$

$$p_{n+k} = \frac{n^k}{n!} p_0 - \ (k-l) p_{n+s} = \frac{n^n}{n!} \frac{s-k}{s-l} p_0 \qquad (k = \overline{l+1, s-1})$$

Етап 2. Розв'язання оптимізаційної задачі пошуку максимального вантаження, що забезпечить виконання умов на допустиму кількість ресурсів обслуговування.

$$\lambda \to max$$

$$\begin{cases} 4 * \left(\sum_{i=1}^{n} i v_k^g \, \mathrm{p}_i + n v_k^g \sum_{i=n+1}^{s} \mathrm{p}_i \right) \le V^g, g = \overline{1, G} \\ \sum_{i=1}^{s} \mathrm{p}_i \le R \end{cases}$$

У білінговій підсистемі балансування вхідного потоку відбувається з використанням модифікованої схеми циклічного розбору (Round Robin), для розрахунку рекомендованого значення інтенсивності вхідного потоку всієї системи буде достатньо проаналізувати метрики з одного DOCS сервера і масштабувати отримані значення на всю підсистему.

Для підстановки вхідних даних буде використана метрика DOCS-OCS-Congestion, зібрана на стороні одного DOCS сервера (Diameter-сервер). В табл. 1 приведено відповідність метрик і лічильників білінгової підсистеми вхідним даним обраної моделі. Всі значення були взяті на підсистемі обслуговування викликів тарифікації інтернет трафіку в момент найбільшої завантаженості ВНТА.

Таблиця 7.1

Відповідність метрик системи вхідним даним моделі

Параметр	**Відповідність метрики**	**Одночасне значення**
n	DOCS-OCS-Congestion	500
μ	DOCS-OCS-Congestion	ВНТА=403/сек
G	DOCS-OCS-Congestion	2048Mb
v_k^g	DOCS-OCS-Congestion	4Mb
V^g	DOCS-OCS-Congestion	1612Mb
s	DOCS-OCS-Congestion	500/сек
R	Визначено експертом	90%
l	DOCS-OCS-Congestion	450

Розрахунок оптимального значення інтенсивності вхідного потоку для існуючої системи

Для розрахунку оптимального значення λ проведено аналіз доступних у білінговій підсистемі метрик для найбільш завантаженого дня в 2017 році (24.11.2017 «Чорна п'ятниця») [124]. Максимальне значення інтенсивності вхідного потоку було розраховано у години найбільшого навантаження, на момент, коли метрики сигналізували про відмови в обслуговуванні (REJECTS) для відомої кількості сесій. Під час аналізу, для кожного запису була розрахована інтенсивність обслуговування заявки (μ) і одночасне значення

інтенсивності вхідного потоку (λ). Базуючись на отриманих даних, можна зробити висновок, що максимальному допустимому значенню інтенсивності вхідного потоку для єдиного OCS процесу, при якому не виникає деградації сервісу (відсутній REJECTS) відповідає λ = 860 при μ = 430. Дані розрахунку оптимального значення λ і μ занесені в табл. 7.2.

Таблиця 7.2

Дані розрахунку оптимального значення λ и μ

DATE	TIME	GPRS1	GPRS2	REJECTS	μ	λ
24.11.2017	00:00	1211112	16102756	0	301	601.93
24.11.2017	01:00	977895	10609138	0	201	401.95
24.11.2017	02:00	962188	9984962	0	190	379.95
24.11.2017	03:00	971583	9719210	0	186	371.96
24.11.2017	04:00	973601	9570451	0	183	365.96
24.11.2017	05:00	1020312	10224994	0	195	389.95
24.11.2017	06:00	1157775	11722724	0	224	447.95
24.11.2017	07:00	1346571	14490450	0	275	549.93
24.11.2017	08:00	1749654	23858101	3	445	889.89
24.11.2017	09:00	1631822	21215782	0	397	793.90
24.11.2017	10:00	1693628	22199421	0	415	829.90
24.11.2017	11:00	1753250	22878271	0	428	855.89
24.11.2017	12:00	1817606	23379187	1	437	873.89
24.11.2017	13:00	1873052	23470265	2	440	879.89
24.11.2017	14:00	1907032	23036244	1	433	865.89
24.11.2017	15:00	1920535	22464995	1	431	861.89
24.11.2017	16:00	1945217	22005529	0	416	831.90
24.11.2017	17:00	1978970	22815321	0	430	859.89
24.11.2017	18:00	1986669	22961832	1	433	865.89
24.11.2017	19:00	1966866	22527160	0	425	849.89
24.11.2017	20:00	1958086	21745802	0	412	823.90
24.11.2017	21:00	1919809	19183564	0	366	731.91
24.11.2017	22:00	1831609	16819098	0	324	647.92
24.11.2017	23:00	1731484	14263415	0	278	555.93

Графічне відображення розрахованої інтенсивності обслуговування заявки (μ) і одночасного значення інтенсивності вхідного потоку (λ) для єдиного OCS процесу зображено на рис. 7.1.

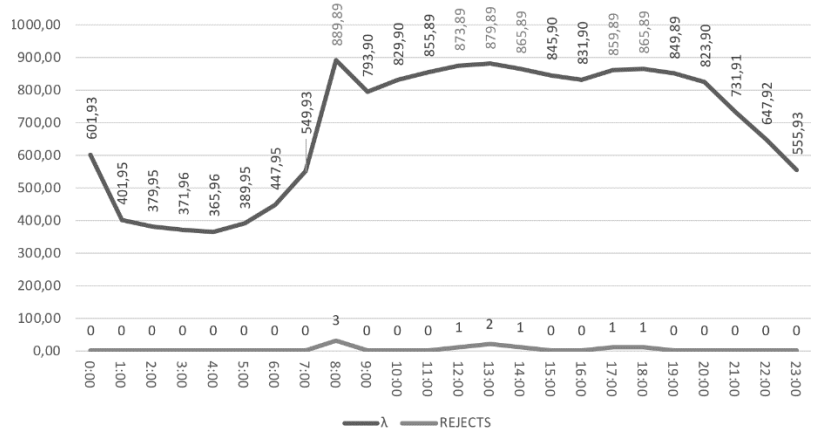


Рис. 7.1 Графік інтенсивності значення λ за 24.11.2017

Розрахунок оптимального значення інтенсивності вхідного потоку для системи, що масштабується

Відповідно до запропонованого розробником системи розширення фізичних компонентів, при масштабуванні системи – кількість OCS процесів, що виконуються на одному DOCS сервері, буде збільшено у 5 разів, при цьому конфігурація OCS процесу змінена не буде. Таким чином, для системи, що масштабується значення констант n, G, v_k^g, Vg, s, R, l, залишаться незмінними, а загальна кількість DOCS серверів буде збільшена.

Базуючись на розрахованому для такої системи значенні λ, можна розрахувати оптимальну кількість OCS процесів необхідних для підсистеми обслуговування викликів тарифікації інтернет трафіку, яка дозволить забезпечити доступність сервісів і цілісність даних, знизить вірогідність втрат і перенавантажень у системі після масштабування.

Враховуючи, що один OCS процес існуючої білінгової підсистеми до масштабування, при оптимальному значенні λ = 860 здатен обробляти трафік з інтенсивністю μ = 430 заявок в секунду, всіх 16 DOCS серверів з одиничним запущеним OCS процесом, оптимальна кількість трафіка, що обробляється буде дорівнювати 6880 сесій за секунду, що відповідає 24 768 000 сесіям в годину, опрацьованим всією підсистемою обслуговування викликів тарифікації інтернет трафіку.

Згідно з планами оператора мобільного зв'язку, інтенсивність вхідного потоку заявок повинна буде бути збільшеною до 50 000 000 сесій за годину. Виходячи з цього розробник білінгової системи дозволив розширити кількість OCS процесів до 30 штук зі збереженням конфігурації, що має забезпечити обробку зростаючого потоку вхідних заявок на білінгову систему. На рис. 7.2 зображено заплановане розширення білінгової підсистеми. Як видно, кількість DOCS серверів було зменшено до 6 штук, при цьому на кожному DOCS сервері запущено по п'ять OCS – процесів, які обробляють логіку дзвінка.

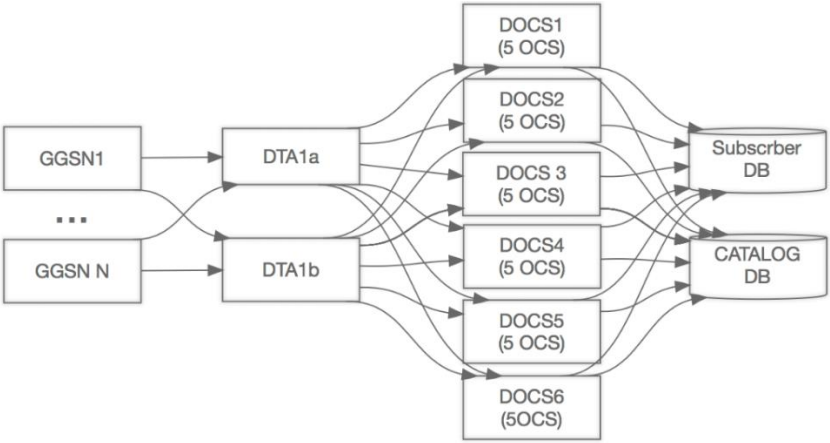


Рис. 7.2 Заплановане розширення білінгової підсистеми

Оскільки, значення інтенсивності вхідного потоку за годину і кількість паралельно працюючих OCS процесів у новій системі відомі, розраховуємо інтенсивність обробки трафіку за секунду, окремо на кожен процес. Використовуючи математичну модель, розраховуємо прогнозоване значення інтенсивності вхідного потоку λ. При 30 одночасно працюючих OCS процесах інтенсивність обробки трафіку за секунду буде μ = 463. У табл. 7.3 відображено вхідні параметри для підставлення в імітаційну модель.

Таблиця 7.3

Вхідні параметри для підставлення в імітаційну модель

Парамет р	Значення
n	500
μ	463/сек
G	2048Mb
v_k^g	4Mb
V^g	1612Mb
s	500/сек
R	90%
l	450

Підставивши ці дані в імітаційну модель, отримаємо **λ = 925.89**, що значно перевищує рекомендоване значення (**λ = 860**).

З вищеописаного слідує, що запропонований варіант масштабування білінгової підсистеми буде недостатнім для задоволення вимог. Більше того, фізична поломка одного DOCS сервера спричинить багатократне збільшення трафіку на DOCS серверах,

які продовжать роботу, що у свою чергу, обов'язково, відобразиться на кількості неопрацьованих замовлень і може призвести до повної відмови всієї білінгової підсистеми.

Керуючись принципом надмірності при архітектурному проектуванні розширення системи, додаємо два DOCS сервера. У такому випадку загальна кількість DOCS серверів, підключених до кластеру, буде дорівнювати 8, а кількість OCS процесів буде збільшено з 30 до 40 штук.

При наявності 40 OCS процесів, інтенсивність обробки трафіку за секунду для кожного OCS процесу буде $\mu = 347$. Після підставлення отриманого значення в імітаційну модель, маємо $\boldsymbol{\lambda = 693.91}$, що задовольняє розраховане рекомендоване значення ($\lambda = 860$).

Навіть у випадку виходу з ладу одного DOCS сервера, прогнозоване значення інтенсивності вхідного потоку не перевищить рекомендоване значення і буде дорівнювати $\boldsymbol{\lambda = 793.90}$, що перевищує відмовостійкість системи. На рис. 7.3 зображено покращене розширення з додаванням двох додаткових DOCS серверів.

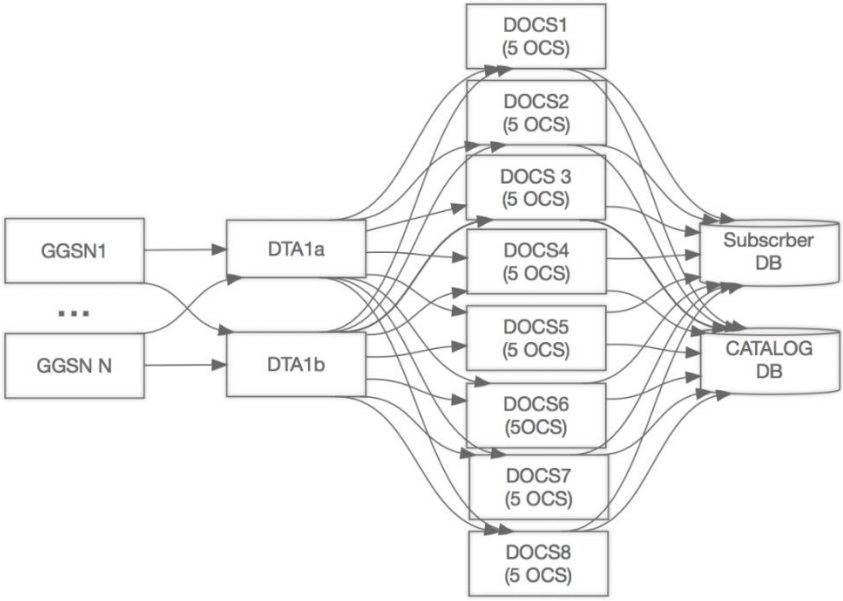


Рис. 7.3 Покращене розширення білінгової підсистеми

Розрахунок номінального значення інтенсивності вхідного потоку після масштабування білінгової системи

Під час проведення робіт по масштабуванню системи були враховані рекомендації по додаванню двох додаткових DOCS серверів. Білінгова. Підсистема була розширена до 8 DOCS серверів. Після переключення трафіку на нову підсистему, недоліків у роботі нової, розширеної підсистеми виявлено не було. Оператором мобільного зв'язку було прийнято рішення збільшити потік вхідних заявок до

запланованої величини – орієнтовно, 50 000 000 заявок за годину. Система успішно витримала навантаження і продовжує роботу у штатному режимі. Через кілька місяців, після збільшення потоку вхідних заявок було проведено порівняльний розрахунок інтенсивності вхідного потоку λ, для підтвердження дієздатності обраного методу. Отримані дані оптимального значення λ і μ, після розширення системи, занесені у табл. 7.4.

Таблиця 7.4

Дані порівняльного розрахунку λ і μ після розширення системи

DATE	TIME	GPRS1	GPRS2	REJECTS	μ	λ
11.04.2018	00:00	2725002	32205511	0	243	485.94
11.04.2018	01:00	2200264	21218275	0	163	325.96
11.04.2018	02:00	2164923	19969924	0	154	07.96
11.04.2018	03:00	2186062	19438419	0	150	299.96
11.04.2018	04:00	2190601	19140901	0	148	295.96
11.04.2018	05:00	2295701	20449988	0	158	315.96
11.04.2018	06:00	2604993	23445448	0	181	361.96
11.04.2018	07:00	3029785	28980900	0	222	443.95
11.04.2018	08:00	3936721	47716201	0	359	717.91
11.04.2018	09:00	3671600	42431565	0	320	639.92
11.04.2018	10:00	3810664	44398842	0	335	669.92
11.04.2018	11:00	3944812	45756542	0	345	689.92
11.04.2018	12:00	4089614	46758374	0	353	705.91
11.04.2018	13:00	4214366	46940530	0	355	709.91
11.04.2018	14:00	4290823	46072488	0	350	699.91
11.04.2018	15:00	4321204	44929989	0	342	683.92
11.04.2018	16:00	4376737	44011058	0	336	671.92
11.04.2018	17:00	4452683	44030643	0	337	673.92
11.04.2018	18:00	4470005	43737265	0	335	669.92
11.04.2018	19:00	4425450	42867923	0	328	655.92
11.04.2018	20:00	4405694	41305206	0	317	633.92
11.04.2018	21:00	4319571	38367128	0	296	591.93
11.04.2018	22:00	4121121	33638196	0	262	523.94
11.04.2018	23:00	3895840	28526830	0	225	449.95

Графічне відображення розрахованої інтенсивності обслуговування заявки (μ) і одночасного значення інтенсивності вхідного потоку (λ), для єдиного OCS процесу, після масштабування системи, зображено на рис. 7.4.

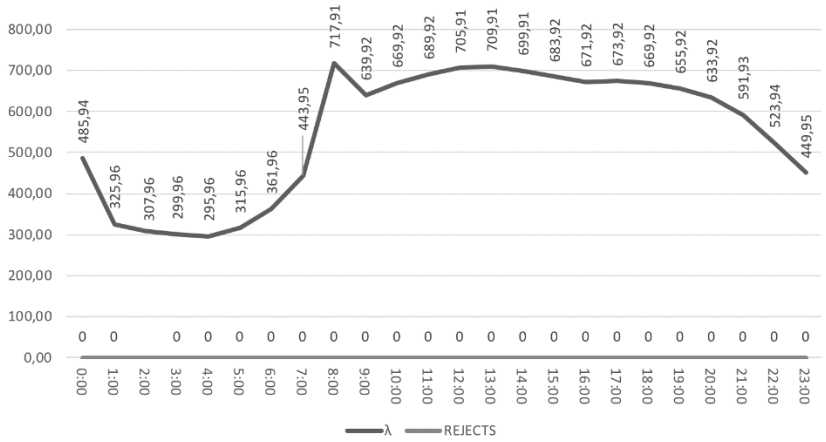


Рис. 7.4 Графік інтенсивності вхідного потоку (λ) за 11.04.2018

Після завершення усіх необхідних розрахунків бачимо, що інтенсивність вхідного потоку не перевищує розраховане рекомендоване значення **λ = 860**, відмови в обслуговуванні – цілком відсутні.

Для повноти аналізу проведемо розрахунок інтенсивності вхідного потоку так, якби рекомендоване додаткове розширення не було проведене, а загальна кількість DOCS серверів залишалась рівною шести.

Отримані дані прогнозованого значення λ і μ, після розширення системи, але з неповною кількістю DOCS серверів, занесені в табл. 7.5.

Графічне відображення прогнозованої інтенсивності обслуговування заявки (μ) и одночасного значення інтенсивності вхідного потоку (λ), для єдиного OCS процесу, після масштабування системи, але без запропонованого додаткового розширення зображено на рис. 7.5.

Таблиця 7.5

Данні порівняльного розрахунку λ і μ після розширення системи, при кількості DOCS серверів рівній 6

DATE	TIME	GPRS1	GPRS2	Передбач. REJECTS	μ	λ
11.04.2018	00:00	2725001,5	32205511	0	323	645.92
11.04.2018	01:00	2200264,3	21218275	0	217	433.95
11.04.2018	02:00	2164922,5	19969924	0	205	409.95
11.04.2018	03:00	2186061,8	19438419	0	200	399.95
11.04.2018	04:00	2190601,4	19140901	0	198	395.95
11.04.2018	05:00	2295701,2	20449988	0	211	421.95
11.04.2018	06:00	2604993,3	23445448	0	241	481.94
11.04.2018	07:00	3029785,2	28980900	0	296	591.93

11.04.2018	08:00	3936721,1	47716201	100	478	955.88
11.04.2018	09:00	3671600,4	42431565	100	427	853.89
11.04.2018	10:00	3810664	44398842	300	446	891.89
11.04.2018	11:00	3944812,3	45756542	300	460	919.89
11.04.2018	12:00	4089614,1	46758374	400	471	941.88
11.04.2018	13:00	4214366	46940530	450	474	947.88
11.04.2018	14:00	4290822,9	46072488	400	466	931.88
11.04.2018	15:00	4321203,9	44929989	300	456	911.89
11.04.2018	16:00	4376737,3	44011058	300	448	895.89
11.04.2018	17:00	4452683,3	44030643	300	449	897.89
11.04.2018	18:00	4470004,5	43737265	300	446	891.89
11.04.2018	19:00	4425449,6	42867923	200	438	875.89
11.04.2018	20:00	4405693,5	41305206	0	423	845.90
11.04.2018	21:00	4319571,1	38367128	0	395	789.90
11.04.2018	22:00	4121120,9	33638196	0	350	699.91
11.04.2018	23:00	3895840	28526830	0	300	599.93

Розрахунок прогнозу показує показує багатократне перевищення розрахованого рекомендованого значення (λ = 860). Також за допомогою методу екстраполяції були спрогнозовані відмови в обслуговуванні (REJECTS).

Після проведеного розширення білінгової підсистеми обслуговування викликів тарифікації, можна зробити висновок, що запропонований метод вибору обчислювальних ресурсів для обслуговування білінгових систем за умови коливання навантаження є оптимальним, початковий вибір цього методу був зроблений вірно.

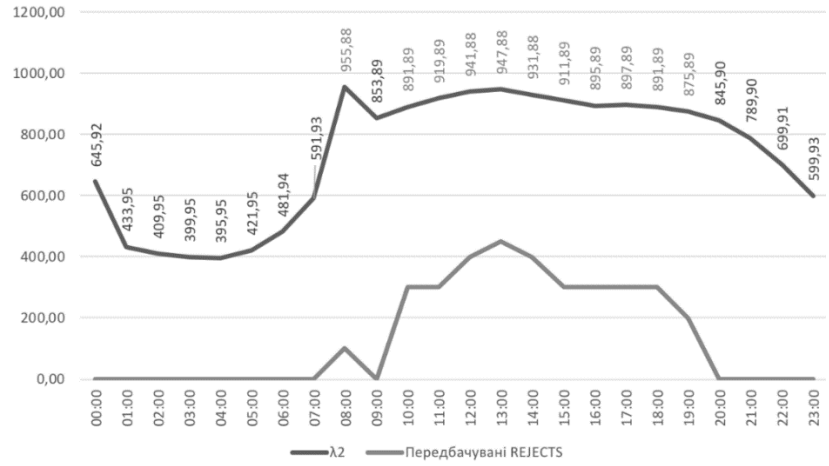


Рис. 7.5 Графік інтенсивності вхідного потоку (λ)

Запропонованій метод і модель оптимального вибору обчислювальних ресурсів, для обслуговування білінгових систем за умови коливання навантаження є ефективними, у повній мірі враховують особливості архітектурних параметрів і можуть бути застосовані у дослідженні для масштабування системи і контролю навантаження білінгової системи.

У ході проведення дослідження масштабування білінгової системи було виявлено, що запропонований варіант масштабування білінгової системи не здатний у повній мірі обробляти запланований до збільшення вхідний потік заявок.

З огляду на виявлений у ході дослідження суттєвий недолік масштабування білінгової системи, було запропоновано і реалізовано додаткове збільшення ємності DOCS серверів, що дозволило забезпечити доступність сервісів і збереження цілісності даних, підвищило відмовостійкість і знизило вірогідність переванатажень системи.

7.2 Імітаційна модель MATLAB методу формування вхідного потоку навантаження для ефективного використання ресурсів обслуговування

Для розв'язання задачі розроблено 4 файли.

1. Файл res.m призначений для розрахунку імовірностей та визначення цільової функції g.

В завданні вигляд цільової функції, максимум якої слід знайти, не визначено. Обрано цільову функцію у вигляді суми всіх ймовірностей, починаючи з p_1. Вибір можна змінити.

```
function    [g]=res()
global  n          %  кількість  каналів  для  одночасного
обслуговування заявок
global  s       %  допустима  кількість  запитів  у  черзі  на
обслуговування
global  l    % кількість запитів у черзі, до досягнення якої
блокується надходження
global p    % масив імовірностей
global alpha beta   %

if beta==1
    A=alpha;
    for k=1:1:n
    A=A+alpha^k/factorial(k);
    end
    p(1)=1/A;
    for k=2:1:(n+1)
        p(k)=n^(k-1)/factorial(k-1)*p(1);
    end
    for k=1:1:l
        p(n+k+1)=n^(n)/factorial(k-1)*p(1);
    end
    for k=(n+l+1):1:(n+s-1)
```

```
            p(n+k+1)=n^(n)/factorial(n)*(s-l)*p(1);
        end
        for k=(l+1):1:(s-1)
            p(n+k+1)=n^(n)/factorial(n)*p(1)-(k-
l)*resm(n^(n)/factorial(n)*(s-l)*p(1));
        end
    else
        A=alpha;
        for k=1:1:n
        A=A+alpha^k/factorial(k);
        end
        A=(1-beta^s)*A+alpha^n/factorial(n)*((beta-beta^(s-
l+1))/(1-beta)-(s-1)*beta^(s+1));
        p(1)=(1-beta^(s-1))/A;
        for k=2:1:(n+1)
            p(k)=alpha^(k-1)/factorial(k-1)*p(1);
            %p(k)=k;
        end
        for k=1:1:l
            p(n+k+1)=alpha^(n)/factorial(n)*p(1);
            %p(n+k+1)=k;
        end
        for k=(l+1):1:(s-1)
            p(n+k+1)=alpha^(n)/factorial(n)*(beta^k-
beta^s)*p(1)/A;
            %p(n+k+1)=k;
        end
        for k=(n+l+1):1:(n+s-1)
            p(n+s+1)=alpha^(n)/factorial(n)*(1-beta)*beta^s/A;
            %p(n+s+1)=k;
        end
        g=sum(p);    % цільова функція
    end
```

2. Файл rosenbrock.m призначений для надання всіх .необхідних констант та ініціалізації робочих масивів. В ньому слід задати (записати в текстовому редакторі) наступні константи:

n – кількість каналів для одночасного обслуговування заявок.

mu – інтенсивність обслуговування заявки,

G – кількість ресурсів залучених для обслуговування заявок,

vg – об'єм g-го ресурсу необхідний для обслуговування у блоці однієї заявки, $g = \overline{1,G}$

Vg – доступний об'єм ресурсу g-го ресурсу який спільно використовується заявками.

s – допустима кількість запитів у черзі на обслуговування.

R – відсоток заявок, які обслуговуються у системі не більше допустимого часу затримки, визначається експертами

l – кількість запитів у черзі, до досягнення якої блокується надходження запитів до системи відповідно до алгоритмів раннього упередження перевантажень.

```
function f = rosenbrock(x)
global n          % кількість каналів для одночасного обслуговування заявок
global mu   % інтенсивність обслуговування заявки
global G          % кількість ресурсів залучених для обслуговування заявок
global vg      % об’єм g-го ресурсу    необхідний для обслуговування у блоці однієї заявки
global Vg    % доступний об’єм ресурсу g-го ресурсу який спільно використовується
global s       % допустима кількість запитів у черзі на обслуговування
global R    % відсоток заявок, які обслуговуються у системі не більше доп.часу
global l    % кількість запитів у черзі, до досягнення якої блок.
global p    % масив імовірностей
global alpha beta   %
global coeffs % масив вагових коефіцієнтів
global y1 y2
%x(1)=lambda  - рекомендоване значення для інтенсивності вхідного потоку

n=3;
y1=zeros(1,n);       % допоміжний масив
y2=zeros(1,s);       % допоміжний масив
coeffs=zeros(1,n); %формування масиву вагових коефіцієнтів
for i=1:1:n
    coeffs(1,i)=i;
end
mu=2;
G=3;
vg=[0.7 0.9  0.9];
Vg=[0.5 0.5 0.5];
s=4;
R=0.6;
l=3;
alpha=x(1)/mu;
beta=alpha/n;
f=res;
```

3. Файл unitdisk.m призначений для формування обмежень на кількість ресурсів обслуговування.

```
function [c, ceq] = unitdisk(x)
global n          % кількість каналів для одночасного обслуговування заявок
global G          % кількість ресурсів залучених для обслуговування заявок
global vg       % об’єм g-го ресурсу   необхідний для обслуговування у блоці однієї заявки
global Vg    % доступний об’єм ресурсу g-го ресурсу який спільно використовується
global s       % допустима кількість запитів у черзі на обслуговування
global R    % відсоток заявок, які обслуговуються у системі не більше доп.часу
global p    % масив імовірностей
global coeffs % масив вагових коефіцієнтів
global y1 y2

for i=1:1:n
    y1(1,i)=p(i+1);
end
for i=1:1:s
    y2(1,i)=p(i+1);
end
yy=y1*coeffs';  % скалярний добуток імовірностей на вагові коефіцієнти
yy2=sum(y2)-R;  % обмеження на суму імовірностей
c=yy*vg-Vg;     % обмеження на ресурси
c(1,n+1) = yy2; % загальний вектор обмежень
ceq = [ ];      % обмеження на точну рівність відсутні
```

4. Файл start.m призначений для запуску програми на виконання. Використовується функція оптимізації fmincon. В перших квадратних дужках необхідно показати початкове значення λ.

```
options  =  optimset('Display','iter','Algorithm','active-set');
[x,fval] = fmincon(@rosenbrock,[0.1],...
                   [],[],[],[],[],[],@unitdisk,options)
```

Для запуску програми слід переписати всі вказані файли в робочу папку Matlab і запустити файл start.m.

Результат розрахунку для вказаного прикладу:

```
     >> start
                                        Max            Line  search
Directional  First-order
      Iter  F-count            f(x)     constraint     steplength
derivative   optimality Procedure
         0       2      10.3813     -0.03288
         1         5        5.52452       -0.01624             0.5
-99.9          75.7
         2         7        4.77695       -0.01031               1
-26.3          19.2
         3        10        4.28726      -0.005139             0.5
-19.2          8.25
         4        12        3.90001    1.411e-005               1
-15.1          1.12
         5        14        3.90096    1.017e-010               1
12.2       0.00115
     Local minimum found that satisfies the constraints.
     Optimization completed because the objective function is
non-decreasing in
     feasible directions, to within the default value of the
function tolerance,
     and constraints were satisfied to within the default value
of the constraint tolerance.
     <stopping criteria details>
     Active inequalities (to within options.TolCon = 1e-006):
       lower      upper     ineqlin   ineqnonlin
                                          2
                                          3
     x =    0.2853
     fval =    3.9010
```

7.3 Імітаційна модель системи онлайн тарифікації з додатковим сервером в середовищі GPSS

В даному розділі запропонована імітаційна модель системи з додатковим сервером в середовищі GPSS, яка ілюструє процес обробки заявок на сервері мобільного оператора враховуючи особливості процесу тарифікації, а також дає інформацію про кількість оброблених та/або втрачених заявок на основному, додатковому сервері та час обслуговування заявок.

За результатами моделювання проаналізована доцільність використання системи обслуговування абонентів мобільного оператора з додатковим сервером.

Вхідні дані

Під час обробки вхідний потік заявок проходить п’ять функціональних блоків. При проходженні кожного з блоків, на обробку заявок виділяється певна кількість

ресурсів – процесорного часу, оперативної пам'яті, постійної пам'яті, зайнятості каналу сигнальним трафіком.

Об'єм займаємих ресурсів для обробки заявок різних сервісів представлений у табл. 7.6.

Таблиця 7.6

Об'єм займаємих ресурсів для обробки заявок різних сервісів

Функціональний блок	Використання ресурсів			Час використання ресурсів, мс
FB1.1	S1	R1	1000	2
		R2	5	
		R3	1000	Depends on NW speed
		R4	0	
	S2	R1	1000	2
		R2	5	
		R3	1000	Depends on NW speed
	S3	R1	1000	2
		R2	5	
		R3	1000	Depends on NW speed
		R4	0	
	S4	R1	1000	2
		R2	5	
		R3	1000	Depends on NW speed
FB1.2	S1	R1	1000	2
		R2	1	
		R3	1000	Depends on NW speed
		R4	0	
	S2	R1	1000	2
		R2	1	
		R3	1000	Depends on NW speed
		R4	0	
	S3	R1	1000	2
		R2	1	
		R3	1000	Depends on NW speed
		R4	0	
	S4	R1	1000	2
		R2	1	
		R3	1000	Depends on NW speed
		R4	0	
FB2.1	S1	R1	0	
		R2	0	
		R3	0	
		R4	0	
	S2	R1	0	

		R2	5	
		R3	1000	
		R4	1000	300
	S3	R1	0	
		R2	6	
		R3	2000	
		R4	2000	3600
FB2.1	S1	R1	0	
		R2	0	
		R3	0	
		R4	0	
	S2	R1	0	
		R2	5	
		R3	1000	
		R4	1000	300
	S3	R1	0	
		R2	6	
		R3	2000	
		R4	2000	3600
	S4	R1	0	
		R2	3	
		R3	500	
		R4	500	1
FB2.2	S1	R1	0	
		R2	0	
		R3	0	
		R4	0	
	S2	R1	1000	300
		R2	5	
		R3	1000	
		R4	0	
	S3	R1	2000	3600
		R2	6	
		R3	2000	
		R4	0	
	S4	R1	500	1
		R2	3	
		R3	500	
		R4	0	
FB3	S1	R1	500	1000
		R2	100	
		R3	500	
		R4	0	

	S2	R1	1000	1000
		R2	200	
		R3	1000	
		R4	0	
	S3	R1	1000	1000
		R2	500	
		R3	1000	
		R4	0	
	S4	R1	600	1000
		R2	200	
		R3	600	
		R4	0	
FB5	S1	R1	200	200
		R2	100	
		R3	200	
		R4	0	
	S2	R1	1000	200
		R2	200	
		R3	1000	
		R4	0	
	S3	R1	2000	200
		R2	600	
		R3	2000	
		R4	0	
	S4	R1	400	200
		R2	150	
		R3	400	
		R4	0	

R1 – об’єм оперативної пам’яті, що використовується,

R2 – Процесорний час

R3 – Занятість каналу сигнальним трафіком

R4 – Об’єм постійної пам’яті

При обробці заявок SMS використовуються функціональні блоки FB1.1, FB1.2, FB2.1, FB2.2, FB3, FB5.

При збільшенні потоку заявок на сервер, ресурсів, що виділяються для обробки потоку може бути не достатньо. Тому заявки можуть втрачатися і якість обслуговування відповідно зменшується.

Якщо при проходженні наступного функціонального блоку кількість втрачених заявок перевищує пороговий рівень (імовірність відказу P>X), то потік заявок направляється на додатковий сервер.

Якщо заявка очікує своєї черги на обслуговування більше заданого часу, то вона також передається на додатковий сервер.

Для визначення кількості необроблених заявок, середній час обробки заявок , коефіцієнти використання каналів обслуговування скористаємось пакетом моделювання систем масового обслуговування GPSS.

Вхідний потік є Пуассонівським. Відомі середній час надходження та обробки заявок. Можливе одночасне надходження декількох заявок.

Середній час обробки заявки вибирається як 2max(τ), де τ – час найдовшого зайняття одного із ресурсів.

Необхідно визначити критерій якості обслуговування, тобто максимальну кількість заявок, які можуть бути не обслужені. При досягненні цього числа, наступні заявки, що надходять на основний сервер будуть передані на додатковий.

Так як, обчислювальні потужності не дозволяють провести моделювання для безкінечної кількості заявок, визначимо кількість заявок, що надходять на основний сервер.

Приймемо кількість заявок, що надходять на основний сервер на протязі часу моделювання – 25000.

Приймемо максимально допустимий час очікування заявки в черзі на обробку перед кожним функціональним блоком – 3 секунди, а на додатковому сервері – 5 секунд.

Аналіз отриманих результатів

Було згенеровано 22486 заявок за законом Пуассона з параметрами (1, 0,001) протягом 30 секунд машинного часу моделювання (рис. 7.6).

LABEL	LOC	BLOCK TYPE	ENTRY COUNT	CURRENT COUNT	RETRY
	1	GENERATE	22486	0	0

Рис. 7.6 Кількість згенерованих заявок

За результатами моделювання втрати на функціональних блоках на основному сервері склали: на першому функціональному блоці – 3250 заявок, на другому – 54, на третьому – 281, на четвертому – 112, на п’ятому 53.

Отже, загальна кількість втрачених заявок при обробці на основному сервері складає 3250+54+281+112+53 = 3750 заявок, що складає 16% усіх заявок.

До першого функціонального блоку додаткового сервера надійшло 10959 заявок, тобто 48% усіх заявок (рис. 7.7).

FBLOK1	54	ENTER	10959	0	0
	55	ENTER	10959	0	0
	56	ADVANCE	10959	49	0

Рис. 7.7 Кількість заявок, що надійшла на додатковий сервер

За результатами моделювання втрати на функціональних блоках на додатковому сервері склали:

На першому функціональному блоці – 49, на другому – 0, на третьому – 25, на четвертому – 426, на п’ятому –0.

Отже, загальна кількість втрачених заявок на додатковому сервері складає

$$49 + 0 + 25 + 426 + 0 = 500,$$

що становить 4,56% від усіх отриманих на додатковому сервері заявок.

Загальний час обробки усіх заявок на додатковому сервері складає сумму часу на обробку в кожному функціональному блоці та час на передачу заявок від основного до додаткового сервера (7 стовпець на рис. 7.8).

$$t_{заг} = 5{,}275 + 1{,}633 + 2{,}508 + 3{,}358 + 11{,}693 + 0{,}2 = 24{,}667 \text{ секунд.}$$

Тоді, в середньому на обробку однієї заявки витрачається 0,0022 секунди що є прийнятним результатом.

QUEUE	MAX	CONT.	ENTRY	ENTRY(0)	AVE.CONT.	AVE.TIME	AVE.(-0)
OCH01	8702	7837	22535	783	3505.090	4.666	4.834
OCH02	501	500	14698	14198	314.697	0.642	18.882
OCH03	699	698	14198	13500	288.980	0.611	12.420
OCH04	432	431	13500	13069	30.397	0.068	2.116
OCH05	1236	1235	12818	11583	582.417	1.363	14.148
PEREH_OCHER	20209	20209	20209	0	9120.827	13.540	13.54
OCHERED01_SERV2	9250	9201	20160	1166	3544.783	5.275	5.599
OCHERED02_SERV2	1461	1460	10910	9450	593.987	1.633	12.205
OCHERED03_SERV2	1500	1499	9450	7951	789.997	2.508	15.810
OCHERED04_SERV2	1971	1970	7926	5956	887.203	3.358	13.511
OCHERED05_SERV2	3944	3944	5530	1586	2155.357	11.693	16.395

Рис. 7.8 Час обробки заявок

Отже, система з додатковим сервером дає низькі показники втрат заявок, а саме 4,56% втрачених заявок, натомість втрати на основному сервері складають близько 16%. Час обробки однієї заявки, враховуючи затримку при передачі на додатковий сервер складає 0,0022с, що є задовільним показником.

7.4 Моделювання задачі розподілу ресурсів між гібридними телекомунікаційними сервісами сервісами

Для розв'язання задачі розподілу ресурсів на сервері оператора мобільного зв'язку необхідно мати такі дані: кількість прибутку, який може бути отриманий від обслуговування однієї заявки кожного типу, загальний об'єм ресурсів, яким розполагає сервер; кількість ресурсів, необхідних усім функціональним блокам для обслуговування заявок кожного типу; статистичні дані про надходження викликів на сервер у різні періоди часу.

Вирішимо задачу розподілу ресурсів між гібридними ТК севісами для системи онлайн тарифікації, де обслуговування чотирьох типів заявок (sms, дзвінки, передача даних через мережу Інтернет, mms) проводиться у п'яти функціональних блоках. Розв'яжемо задачу для двох типів ресурсів: оперативної (R1) і постійної (R2) пам'яті сервера. Для прикладу візьмемо дані про кількість ресурсів, необхідних кожному функціональному блоку для обслуговування заявки певного типу, представлені в табл. 7.7.

Таблиця 7.7

Розподіл ресурсів між функціональними блоками

№ ФБ	FB1.	FB1.2	FB2.	FB2.2	FB3	FB4.1	FB4.2	FB5

Resource usage	R1 (байти)	S1	1000	1000	0	0	500	1000	1000	200
		S2	1000	1000	0	1000	1000	1000	100	1000
		S3	1000	1000	0	2000	1000	1000	100	2000
		S4	1000	1000	0	500	600	1000	100	400
	R2 (байти)	S1	0	0	0	0	0	100	1000	0
		S2	0	0	1000	0	0	100	1000	0
		S3	0	0	2000	0	0	100	1000	0
		S4	0	0	500	0	0	100	1000	0

S1 – дзвінки; S2 – sms-повідомлення; S3 – mms-повідомлення; S4 – Інтернет-контент.

В табл. 7.8 представлені основні задачі, які виконують функціональні блоки. У першому, другому і четвертому функціональних блоках заявки можуть обслуговуватись двома способами. При цьому кількість ресурсів залежатиме від того, який зі способів був обраний.

Таблиця 7.8

Призначення функціональних блоків

Функціональний блок	Задачі
FB1.1	Gathering subscriber data from Permanent Storage
FB1.2	Gathering subscriber data from cache
FB2.1	State write/read to permanent storage
FB2.2	State write/read to cache
FB3	Rating
FB4.1	Online notification
FB4.2	Offline notification
FB5	Final debit

У табл. 7.9 представлена кількість прибутку (коефіцієнт корисності для оператора), яку може отримати оператор від обслуговування однієї заявки кожного типу сервісу в умовних одиницях.

Таблиця 7.9

Прибуток від кожного сервісу

Сервіс	Величина прибутку
Sms	5 у.о.
Voice	2 у.о.
Data transfer	6 у.о.
mms	3 у.о.

Прибуток від кожного сервісу залежить від тарифів у даний момент часу, які не є фіксованими, тому дані в таблиці взяті для прикладу.

Для того, щоб розв'язати задачу розподілу ресурсів на сервері оператора мобільного зв'язку, необхідно ввести додаткові обмеження для кількості заявок кожного типу. Для цього потрібно аналізувати статистичні дані про кількість заявок в кожний період часу.

Для прикладу розглянемо співвідношення між кількістю заявок різних типів у період із 12 до 13 години (рис. 7.9)

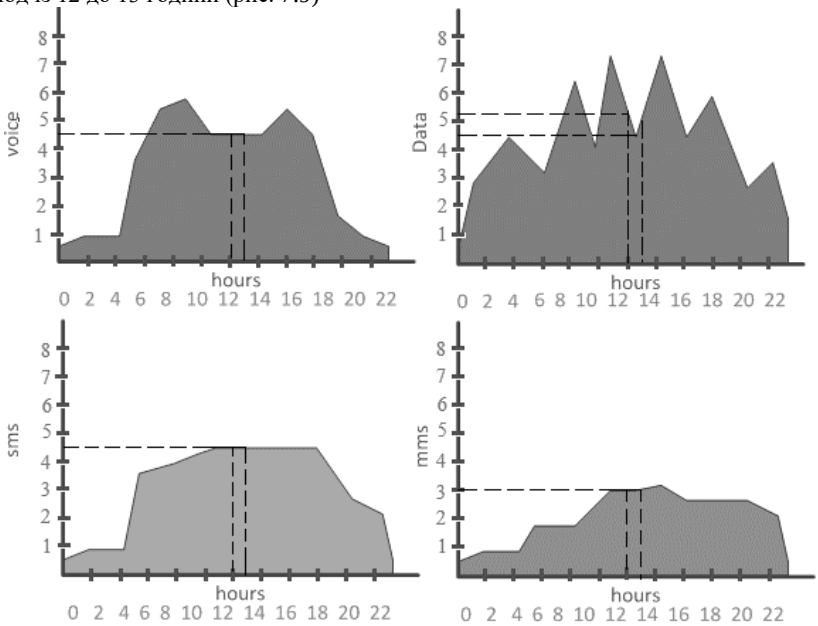


Рис. 7.9 Графіки навантаження по кожному сервісу

З рисунка видно, що кількість дзвінків, sms, mms та Інтернет-контенту у цей період складає в середньому 4,5; 4,5; 3 і 4,8 одиниць відповідно. Для того, щоб отримати обмеження на кількість заявок кожного типу, знайдемо співвідношення між ними у даний період часу.

Отримаємо такі співвідношення:

$$\frac{S1}{S2} = \frac{4,5}{4,5} = 1$$
$$\frac{S1}{S3} = \frac{4,5}{3} = 1,5$$
$$\frac{S1}{S4} = \frac{4,5}{4,8} = 0,94$$
$$\frac{S2}{S3} = \frac{4,5}{3} = 1,5$$
$$\frac{S2}{S4} = \frac{4,5}{4,8} = 0,94$$
$$\frac{S3}{S4} = \frac{3}{4,8} = 0,625$$

Розширивши діапазон значень від -0,5 до 0,5, отримаємо такі додаткові обмеження: $0,5 \leq \frac{S1}{S2} \leq 1,5$

$$1 \leq \frac{S1}{S3} \leq 2$$
$$0,44 \leq \frac{S1}{S4} \leq 1,44$$
$$1 \leq \frac{S2}{S3} \leq 2$$
$$0,44 \leq \frac{S2}{S4} \leq 1,44$$
$$0,125 \leq \frac{S3}{S4} \leq 1,125$$

Перейдемо до вирішення задачі розподілу ресурсів між сервісами з урахуванням економічної ефективності обслуговування заявок.

Математичне описання задачі наступне:

$$\sum_{i=1}^{4} k_i S_i \rightarrow max$$

$$\begin{cases} \sum_{i=1}^{4} k_i (\sum_{j=1}^{8} v_{ij}^{R1}) \le V_{R1} \\ \sum_{i=1}^{4} k_i (\sum_{j=1}^{8} v_{ij}^{R2}) \le V_{R2} \\ 0{,}5 \le \frac{k_1}{k_2} \le 1{,}5 \\ 1 \le \frac{k_1}{k_3} \le 2 \\ 0{,}44 \le \frac{k_1}{k_4} \le 1{,}44 \\ 1 \le \frac{k_2}{k_3} \le 2 \\ 0{,}44 \le \frac{k_2}{k_4} \le 1{,}44 \\ 0{,}125 \le \frac{k_3}{k_4} \le 1{,}125 \end{cases}$$

Де k_1 - кількість заявок 1-го типу (дзвінки);

k_2 - кількість заявок 2-го типу (sms-повідомлення);

k_3 – кількість зявок 3-го типу (mms-повідомлення);

k_4 – кількість зявок 4-го типу (Інтернет-контент);

S_i – кількість прибутку, отриманого від обслуговування однієї заявки певного типу.

З таблиці видно, що є два варіанти проходження заявки через функціональні блоки, а саме через 1-й, 2-й та 4-й. Тому при вирішенні поставленої задачі будемо розглядати обидва способи.

Щоб визначити величини V_{R1} і V_{R2} (приблизну кількість оперативної і постійної пам'яті, яку може надати сервер), припустимо, що на сервер надходить по 10^6 заявок кожного типу. Скориставшись відомою інформацією про те, яку кількість ресурсів потребує кожен тип заявок у кожному функціональному блоці, підрахуємо загальну кількість оперативної і постійної пам'яті, необхідної для обслуговування усіх заявок.

Розглянемо випадок, коли заявка проходить наступний шлях:

FB1.1→FB2.1→FB3→FB4.1→FB5

1) Загальна кількість оперативної пам'яті

Кількість оперативної пам'яті, необхідної для обслуговування 1 заявки кожного типу.

S1: $V_{R1}{}^1 = 1000 + 0 + 500 + 1000 + 200 = 2700$

S2: $V_{R1}{}^2 = 1000 + 0 + 1000 + 1000 + 1000 = 4000$

S3: $V_{R1}{}^3 = 1000 + 0 + 1000 + 1000 + 2000 = 5000$

S4: $V_{R1}{}^4 = 1000 + 0 + 600 + 1000 + 400 = 3000$

$V_{R1} = 10^6 \cdot (V_{R1}^1 + V_{R1}^2 + V_{R1}^3 + V_{R1}^4) = 10^6 \cdot (2700 + 4000 + 5000 + 3000) =$

$= 1{,}47 \cdot 10^{10}$

2) Загальна кількість постійної пам'яті

Кількість постійної пам'яті, необхідної для обслуговування 1 заявки кожного типу.

S1: ${V_{R2}}^1 = 0 + 0 + 0 + 100 + 0 = 100$

S2: ${V_{R2}}^2 = 0 + 1000 + 0 + 100 + 0 = 1100$

S3: ${V_{R2}}^3 = 0 + 2000 + 0 + 100 + 0 = 2100$

S4: ${V_{R2}}^4 = 0 + 500 + 0 + 100 + 0 = 600$

$V_{R2} = 10^6 \cdot (V_{R2}^1 + V_{R2}^2 + V_{R2}^3 + V_{R2}^4) = 10^6 \cdot (100 + 1100 + 2100 + 600) =$

$= 3{,}9 \cdot 10^9$

Тобто загальна кількість оперативної і постійної пам'яті, яку можуть використовувати функціональні блоки для обслуговування заявок не може перевищувати ці числа.

Для вирішення задачі оптимізації цільової функції використовується пакет Microsoft Office (Excel). Щоб знайти параметри k_i (i=1,4), при яких цільова функція буде максимальною, скористаємося інструментом «Пошук рішення» на панелі інструментів Excel (рис. 7.10).

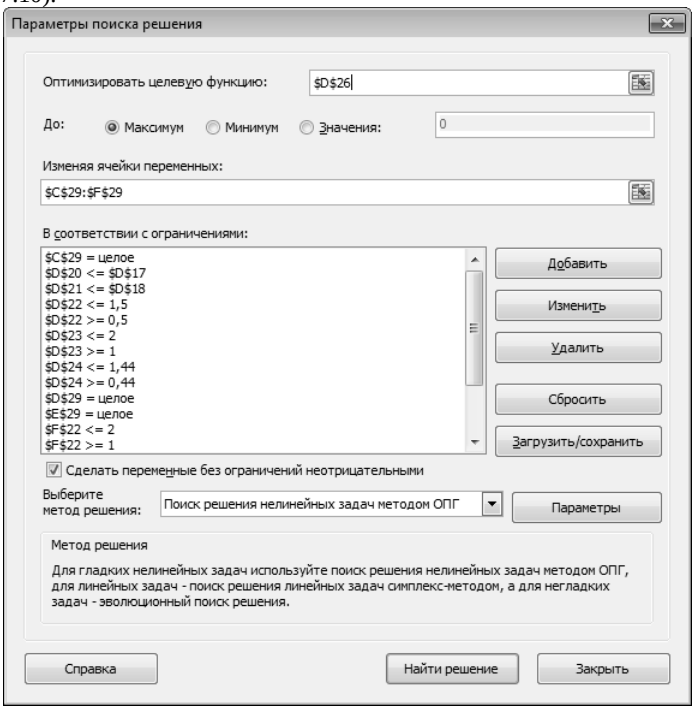


Рис. 7.10 Панель інструментів «Пошук рішення» MS Excel

Задаємо номер комірки з цільовою функцією, яку необхідно оптимізувати до максимуму і задаємо визначені вище обмеження. Крім того, необхідно вказати, що знайдені значення кількості заявок повинні бути цілими числами. Після визначення усіх необхідних умов знаходимо рішення задачі.

В результаті отримаємо такі значення кількості заявок:

Таблиця 7.10

Кількість заявок кожного типу

k1	k2	k3	k4
1247676	831784	831784	1281739

При цьому значення цільової функції, тобто максимального прибутку від обслуговування усіх заявок буде наступним:

$$\sum_{i=1}^{4} k_i S_i = 16737869(\text{у. о.})$$

Перевіримо виконання усіх необхідних умов.

1) $\sum_{i=1}^{4} k_i(\sum_{j=1}^{8} v_{ij}^{R1}) = 1247676 \cdot (1000 + 0 + 500 + 1000 + 200) + 831784 \cdot (0 + 1000 + 0 + 100 + 0) + 831784 \cdot (1000 + 0 + 1000 + 1000 + 2000) + 1281739 \cdot (1000 + 0 + 600 + 1000 + 400) = 1{,}47 \cdot 10^{10} \leq V_{R1}$;

2) $\sum_{i=1}^{4} k_i(\sum_{j=1}^{8} v_{ij}^{R2}) = 1247676 \cdot (0 + 0 + 0 + 100 + 0) + 831784 \cdot (1000 + 0 + 1000 + 1000 + 1000) + 831784 \cdot (0 + 2000 + 0 + 100 + 0) + 1281739 \cdot (0 + 500 + 0 + 100 + 0) = 3{,}56 \cdot 10^{9} \leq V_{R2}$;

3) $\frac{k_1}{k_2} = \frac{1247676}{831784} = 1{,}5$ $\quad 0{,}5 \leq 1{,}5 \leq 1{,}5$;

4) $\frac{k_1}{k_3} = \frac{1247676}{831784} = 1{,}5$ $\quad 1 \leq 1{,}5 \leq 2$;

5) $\frac{k_1}{k_4} = \frac{1247676}{1281739} = 0{,}973424$ $\quad 0{,}44 \leq 0{,}973424 \leq 1{,}44$;

6) $\frac{k_2}{k_3} = \frac{831784}{831784} = 1$ $\quad 1 \leq 1 \leq 2$;

7) $\frac{k_2}{k_4} = \frac{831784}{1281739} = 0{,}64895$ $\quad 0{,}44 \leq 0{,}64895 \leq 1{,}44$;

8) $\frac{k_3}{k_4} = \frac{831784}{1281739} = 0{,}64895$ $\quad 0{,}125 \leq 0{,}64895 \leq 1{,}125$;

Тобто усі необхідні умови виконуються при даних значеннях k_i.

Визначимо необхідний об'єм ресурсів для знайденої кількості заявок кожного типу за формулами:

$V_i^{R1} = k_i \sum_j v_{ij}^{R1}$ – кількість оперативної пам'яті;

$V_i^{R2} = k_i \sum_j v_{ij}^{R2}$ – кількість постійної пам'яті.

Отримаємо такі результати:

Таблиця 7.11

Розподіл ресурсів між заявками

R1	S1	S2	S3	S4
	3,37E + 09	3,33E + 09	4,16E + 09	3,8E + 09
R2	S1	S2	S3	S4

	1,25E + 08	9,15E + 08	1,75E + 09	7,69E + 08

Тобто кількість оперативної пам'яті для кожного типу сервісу дорівнює:

$V_1^{R1} = 3{,}37 \cdot 10^9$ - для дзвінків;

$V_2^{R1} = 3{,}33 \cdot 10^9$ – для sms-повідомлень;

$V_3^{R1} = 4{,}16 \cdot 10^9$ – для mms-повідомлень;

$V_4^{R1} = 3{,}8 \cdot 10^9$ – для Інтернет-контенту.

Кількість постійної пам'яті для кожного типу сервісу дорівнює:

$V_1^{R2} = 1{,}25 \cdot 10^8$ - для дзвінків;

$V_2^{R2} = 9{,}15 \cdot 10^8$ – для sms-повідомлень;

$V_3^{R2} = 1{,}75 \cdot 10^9$ – для mms-повідомлень;

$V_4^{R2} = 7{,}69 \cdot 10^8$ – для Інтернет-контенту.

Вирішимо аналогічну задачу для випадку, коли заявка проходить через інші функціональні блоки:

FB1.2→FB2.2→FB3→FB4.2→FB5

Кількість оперативної пам'яті, необхідної для обслуговування 1 заявки кожного типу.

S1: ${V_{R1}}^1 = 1000 + 0 + 500 + 1000 + 200 = 2700$

S2: ${V_{R1}}^2 = 1000 + 1000 + 1000 + 100 + 1000 = 4100$

S3: ${V_{R1}}^3 = 1000 + 2000 + 1000 + 100 + 2000 = 6100$

S4: ${V_{R1}}^4 = 1000 + 500 + 600 + 100 + 400 = 2600$

$V_{R1} = 10^6 \cdot (V_{R1}^1 + V_{R1}^2 + V_{R1}^3 + V_{R1}^4 = 10^6 \cdot (2700 + 4100 + 6100 + 2600) =$
$= 1{,}55 \cdot 10^{10}$

3) Загальна кількість постійної пам'яті

Кількість постійної пам'яті, необхідної для обслуговування 1 заявки кожного типу.

S1: ${V_{R2}}^1 = 0 + 0 + 0 + 1000 + 0 = 1000$

S2: ${V_{R2}}^2 = 0 + 0 + 0 + 1000 + 0 = 1000$

S3: ${V_{R2}}^3 = 0 + 0 + 0 + 1000 + 0 = 1000$

S4: ${V_{R2}}^4 = 0 + 0 + 0 + 1000 + 0 = 1000$

$V_{R2} = 10^6 \cdot (V_{R2}^1 + V_{R2}^2 + V_{R2}^3 + V_{R2}^4) = 10^6 \cdot (1000 + 1000 + 1000 + 1000) =$
$= 4 \cdot 10^9$

Таблиця 7.12

Кількість заявок кожного типу

k1	k2	k3	k4
1321102	880735	880731	917432

При цьому значення цільової функції, тобто максимального прибутку від обслуговування усіх заявок буде наступним:

$$\sum_{i=1}^{4} k_i S_i = 16403662(\text{у. о.})$$

Перевіримо виконання усіх необхідних умов.

1) $\sum_{i=1}^{4} k_i(\sum_{j=1}^{8} v_{ij}^{R1}) = 1321102 \cdot (1000 + 0 + 500 + 1000 + 200) + 880735 \cdot (1000 + 1000 + 1000 + 100 + 1000) + 880731 \cdot (1000 + 2000 + 1000 + 100 + 2000) + 917432 \cdot (1000 + 500 + 600 + 100 + 400) = 1{,}49 \cdot 10^{10} \leq V_{R1}$;

2) $\sum_{i=1}^{4} k_i(\sum_{j=1}^{8} v_{ij}^{R2}) = 1321102 \cdot (0 + 0 + 0 + 1000 + 0) + 880735 \cdot (0 + 0 + 0 + 1000 + 0) + 880731 \cdot (0 + 0 + 0 + 1000 + 0) + 917432 \cdot (0 + 0 + 0 + 1000 + 0) = 4 \cdot 10^9 \leq V_{R2}$;

3) $\frac{k_1}{k_2} = \frac{1321102}{880735} = 1{,}499999$ $0{,}5 \leq 1{,}499999 \leq 1{,}5$;

4) $\frac{k_1}{k_3} = \frac{1321102}{880731} = 1{,}500006$ $1 \leq 1{,}500006 \leq 2$;

5) $\frac{k_1}{k_4} = \frac{1321102}{917432} = 1{,}44$ $0{,}44 \leq 1{,}44 \leq 1{,}44$;

6) $\frac{k_2}{k_3} = \frac{880735}{880731} = 1{,}0000045$ $1 \leq 1{,}0000045 \leq 2$;

7) $\frac{k_2}{k_4} = \frac{880735}{917432} = 0{,}9600003$ $0{,}44 \leq 0{,}9600003 \leq 1{,}44$;

8) $\frac{k_3}{k_4} = \frac{880731}{917432} = 0{,}9599959$ $0{,}125 \leq 0{,}9599959 \leq 1{,}125$;

Тобто усі необхідні умови виконуються при даних значеннях k_i.

Визначимо необхідний об'єм ресурсів для знайденої кількості заявок кожного типу за формулами:

$V_i^{R1} = k_i \sum_j v_{ij}^{R1}$ – кількість оперативної пам'яті;

$V_i^{R2} = k_i \sum_j v_{ij}^{R2}$ – кількість постійної пам'яті.

Отримаємо такі результати:

Таблиця 7.13

Розподіл ресурсів між заявками

R1	S1	S2	S3	S4
	3,57E + 09	3,6E + 09	5,4E + 09	2,38E + 09
R2	S1	S2	S3	S4
	1,32E + 09	8,8E + 08	8,8E + 08	9,17E + 08

Тобто кількість оперативної пам'яті для кожного типу сервісу дорівнює:

$V_1^{R1} = 3{,}57 \cdot 10^9$ - для дзвінків;

$V_2^{R1} = 3{,}6 \cdot 10^9$ – для sms-повідомлень;

$V_3^{R1} = 5{,}4 \cdot 10^9$ – для mms-повідомлень;

$V_4^{R1} = 2{,}38 \cdot 10^9$ – для Інтернет-контенту.

Кількість постійної пам'яті для кожного типу сервісу дорівнює:

$V_1^{R2} = 1{,}32 \cdot 10^9$ - для дзвінків;

$V_2^{R2} = 8{,}8 \cdot 10^8$ – для sms-повідомлень;

$V_3^{R2} = 8{,}8 \cdot 10^8$ – для mms-повідомлень;

$V_4^{R2} = 9{,}17 \cdot 10^8$ – для Інтернет-контенту.

Для того, щоб збільшити кількість заявок, які можуть обслуговуватись одночасно, потрібно збільшити кількість відповідних ресурсів, необхідних для їх обробки функціональними блоками. Для цього можна використовувати додаткові віртуальні

сервери у хмарах. Використання хмарних технологій дозволить збільшити об'єм необхідних ресурсів на сервері до будь-якого розміру.

Розглянемо на діаграмах, як буде змінюватись кількість заявок та відповідних ресурсів, необхідних для їх обслуговування при поступовому збільшенні об'єму ресурсів на сервері.

Будемо проводити дослідження залежності кількості заявок чотирьох різних типів та об'єму ресурсів для їх обслуговування, збільшуючи максимальний об'єм ресурсів оперативної та постійної пам'яті, який може бути доступним на сервері. Щоб визначити приріст ресурсів, який буде задаватись у процесі дослідження, визначимо необхідний об'єм оперативної та постійної пам'яті для потрібний функціональний блокам для обслуговування викликів, якщо на вхід сервера надходить два, три, чотири або п'ять мільйонів заявок кожного типу.

1) Для випадку FB1.1→FB2.1→FB3→FB4.1→FB5

Якщо k_i=2000000

$$V_{R1} = 2 \cdot 10^6 \cdot (V_{R1}^1 + V_{R1}^2 + V_{R1}^3 + V_{R1}^4) = 2{,}94 \cdot 10^{10}$$
$$V_{R2} = 2 \cdot 10^6 \cdot (V_{R2}^1 + V_{R2}^2 + V_{R2}^3 + V_{R2}^4) = 7{,}8 \cdot 10^9$$

Якщо k_i=3000000

$$V_{R1} = 3 \cdot 10^6 \cdot (V_{R1}^1 + V_{R1}^2 + V_{R1}^3 + V_{R1}^4) = 4{,}41 \cdot 10^{10}$$
$$V_{R2} = 3 \cdot 10^6 \cdot (V_{R2}^1 + V_{R2}^2 + V_{R2}^3 + V_{R2}^4) = 1{,}17 \cdot 10^{10}$$

Якщо k_i=4000000

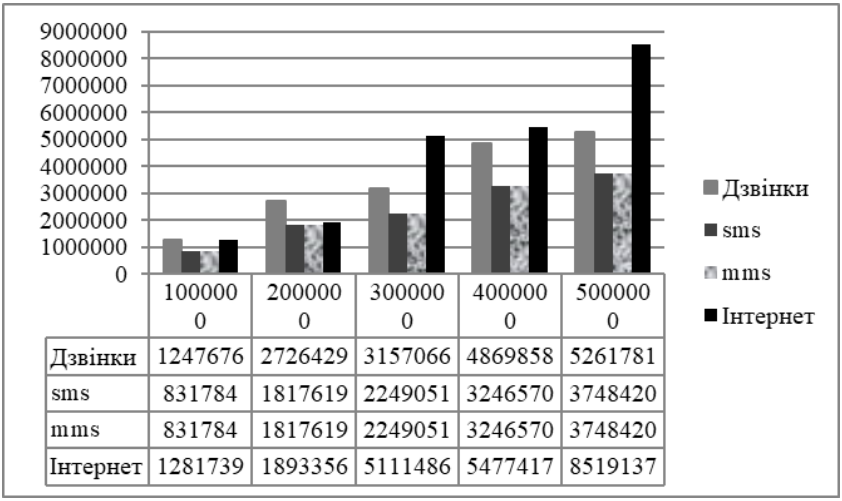
$$V_{R1} = 4 \cdot 10^6 \cdot (V_{R1}^1 + V_{R1}^2 + V_{R1}^3 + V_{R1}^4) = 5{,}88 \cdot 10^{10}$$
$$V_{R2} = 4 \cdot 10^6 \cdot (V_{R2}^1 + V_{R2}^2 + V_{R2}^3 + V_{R2}^4) = 1{,}56 \cdot 10^{10}$$

Якщо k_i=4000000

$$V_{R1} = 5 \cdot 10^6 \cdot (V_{R1}^1 + V_{R1}^2 + V_{R1}^3 + V_{R1}^4) = 7{,}35 \cdot 10^{10}$$
$$V_{R2} = 5 \cdot 10^6 \cdot (V_{R2}^1 + V_{R2}^2 + V_{R2}^3 + V_{R2}^4) = 1{,}95 \cdot 10^{10}$$

Розрахуємо оптимальну кількість заявок кожного типу та об'єм ресурсів, необхідний для їх обслуговування для даних об'ємів оперативної та постійної пам'яті на сервері і зобразимо залежності у вигляді діаграм.

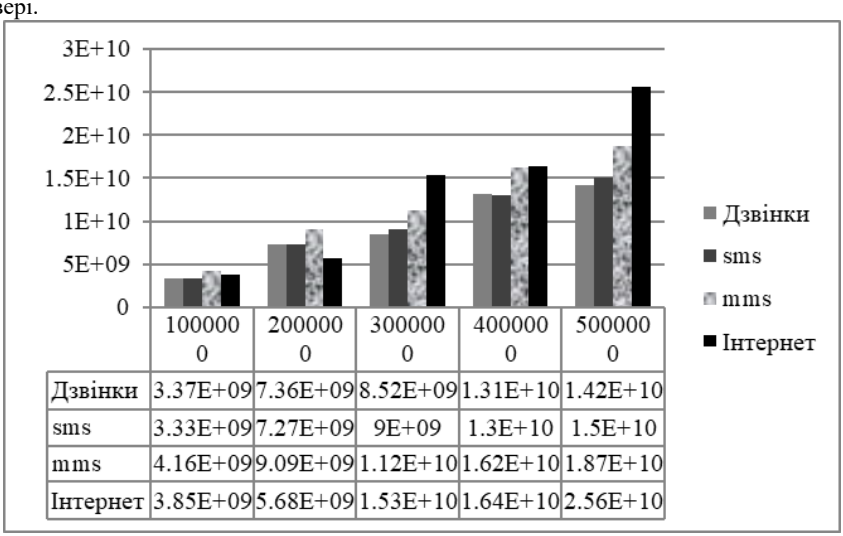
1) Залежність оптимальної кількості заявок чотирьох типів від збільшення максимальної кількості ресурсів сервера.



	1000000	2000000	3000000	4000000	5000000
Дзвінки	1247676	2726429	3157066	4869858	5261781
sms	831784	1817619	2249051	3246570	3748420
mms	831784	1817619	2249051	3246570	3748420
Інтернет	1281739	1893356	5111486	5477417	8519137

Рис. 7.11 Залежність кількості заявок від продуктивності сервера

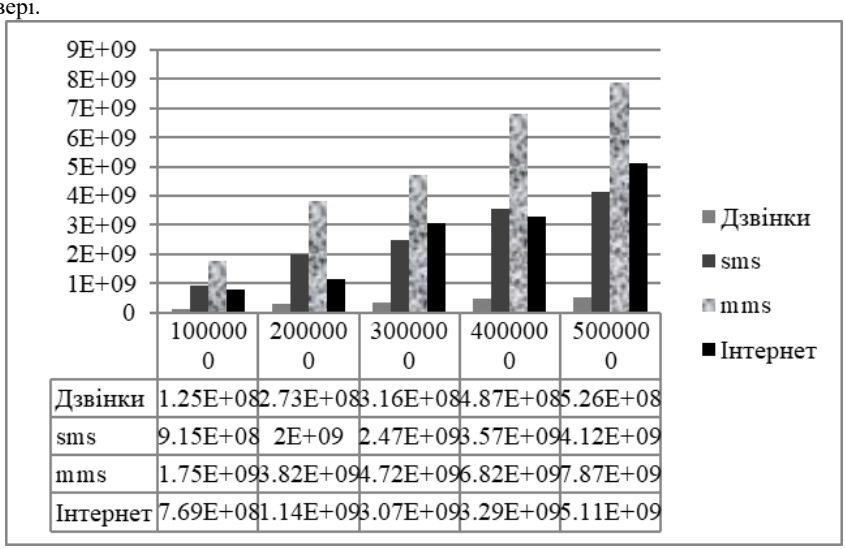
2) Залежність об'єму оперативної пам'яті, який необхідно виділити для обслуговування кожного виду заявок від збільшення загальної кількості ресурсів на сервері.



	1000000	2000000	3000000	4000000	5000000
Дзвінки	3.37E+09	7.36E+09	8.52E+09	1.31E+10	1.42E+10
sms	3.33E+09	7.27E+09	9E+09	1.3E+10	1.5E+10
mms	4.16E+09	9.09E+09	1.12E+10	1.62E+10	1.87E+10
Інтернет	3.85E+09	5.68E+09	1.53E+10	1.64E+10	2.56E+10

Рис. 7.12 Залежність кількості оперативної пам'яті для обслуговування кожного типу заявок від продуктивності сервера

3) Залежність об’єму постійної пам’яті, який необхідно виділити для обслуговування кожного виду заявок від збільшення загальної кількості ресурсів на сервері.



	1000000	2000000	3000000	4000000	5000000
Дзвінки	1.25E+08	2.73E+08	3.16E+08	4.87E+08	5.26E+08
sms	9.15E+08	2E+09	2.47E+09	3.57E+09	4.12E+09
mms	1.75E+09	3.82E+09	4.72E+09	6.82E+09	7.87E+09
Інтернет	7.69E+08	1.14E+09	3.07E+09	3.29E+09	5.11E+09

Рис. 7.13 Залежність кількості постійної пам’яті для обслуговування кожного типу заявок від продуктивності сервера

7.5 Імітаційна модель системи онлайн тарифікації із змінним ресурсом обслуговування

Імітаційна модель програмно-керованої мережі обчислювальних вузлів метою якої є створення інформаційно-обчислювальної мережі, яка динамічно організується на вимогу та виступає технічним засобом для вирішення задач системи тарифікації.

Розробити програмне забезпечення на основі запропонованих методів, що дозволяє розгорнути інформаційну мережу для забезпечення функціонування систем online тарифікації

Будується обчислювальна мережа, де

Вузол мережі – обчислювальна машина (віртуальна машина, яка створюється тільки для виконання однієї або групи задач системи online тарифікації (OCS).

Потоки мережі – інформаційні потоки, що генеруються в процесі обслуговування заявок на тарифікацію; керування потоками здійснюється централізовано за таблицями.

Таблиці керування потоками залежить від структури мережі, яка змінюється відповідно до розкладу.

Розклад включення серверів – залежить від дня року, окремо будується розклад для робочих та святкових днів.

Для не святкових днів розклад переключення записано в матриці M1 розмірністю (NxS), де N – кількість серверів, які планується підключати до системи в процесі

обслуговування абонентів; S=24*7*k – кількість переключень (зміна кількості серверів – учасників комутаційної мережі) за дубу та протягом тижня; k – кількість можливих переключень протягом години у несвятковий день.

Для святкових днів розклад переключення серверів записано в матрицю M2 розмірністю (NxS1), де N – кількість серверів, які планується підключати до системи в процесі обслуговування абонентів; S=24*k1 – кількість переключень (зміна кількості серверів – учасників комутаційної мережі) за дубу; k1 – кількість можливих переключень протягом години у святковий день.

Ресурси вузла мережі – до них відноситься кількість оперативної пам'яті, кількість постійної пам'яті на дисках, потужність процесору , а також вхідний та вихідний катали зв'язку. Ресурси вузла можуть бути змінені динамічно за розкладом розрахованим выдповыдно до запропонованого методу, моделювання методу наведено в п. 7.7.

Оцінювання якості обслуговування виконувалося на наборах статистичних даних, отриманих від компанії оператору зв'язку. У Додатку 1 наведено код програми, яка імітує обслуговування запитів, що надходять на в вхід системи обслуговування, яка складається з набору вузлів (функціональних блоків). Ресурси кожного вузла визначаються конфігураціями та розкладом використання конфігурацій. На вхід подавалася послідовність запитів для обслуговування, послідовність містила $T_{мод}$ наборів запитів, кількість якийх визначалася відповідно до отриманих від оператора зв'язку статистичних даних. На рис. 7.14 наведено схему імітаційної моделі.

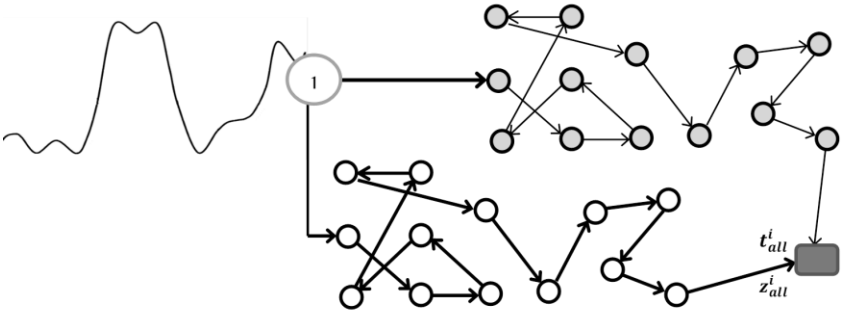


Рис. 7.14 Імітаційна модель ГТС

Вхідні дані імітаційної моделі:

dt = 0,1 сек – інтервал дискретизації часу;

$T_{мод}$ =864000 інтервалів – загальний час моделювання (дискретний);

N^i_Σ – кількість запитів, які надійшли в систему за час i-го моделювання, $i = \overline{1,100}$;

$M = \begin{cases} 28, \text{при викор. додаткових ресурсів} \\ 14, \text{ інакше} \end{cases}$;

$Vj = \{V_{R1j}, V_{R2j}\}$ - доступні об'єми ресурсів вузла j.

Для кожного моделювання була зафіксована статистика показників якості обслуговування для кожного *j-го* вузла, а також в цілому по системі:

t_j^i – середній час затримки у вузлі j ($j = \overline{1,M}$), під час і-го моделювання $i = \overline{1,100}$;

z_j^i – кількість втрачених запитів у вузлі j ($j = \overline{1,M}$), під час і-го моделювання $i = \overline{1,100}$;

t_{all}^i - середній час затримки у системі під час і-го моделювання $i = \overline{1,100}$;

z_{all}^i кількість втрачених запитів у системі під час і-го моделювання $i = \overline{1,100}$.

Також системою моніторингу зберігалися дані про кількість ресурсів у кожний малий інтервал часу:

$R_{1j}{}^i = \{R_{1j}^1, \dots R_{1j}^k, \dots R_{1j}^{T_{мод}}\}$ – множина даних моніторинг ресурсу R1 у вузлі j ($j = \overline{1,M}$), під час і-го моделювання $i = \overline{1,100}$;

$R_{2j}{}^i = \{R_{2j}^1, \dots R_{2j}^k, \dots R_{2j}^{T_{мод}}\}$ - – моніторинг ресурсу R2 у вузлі j ($j = \overline{1,M}$), під час і-го моделювання $i = \overline{1,100}$.

Для оцінки показників якості обслуговування гібридних сервісів виконувався розрахунок ймовірності того, що будуть порушуватися вимоги стандартів та специфікацій щодо для часу обслуговування та ймовірності вчасного обслуговування сервісу, відповідні формули для розрахунку ймовірностей було наведено у табл. 7.14. Результати моделювання зведені у табл. 7.15.

В роботі оператору зв'язку важливим показником функціонування системи в цілому є коефіцієнт використання ресурсів. Практика телекомунікаційної компанії показала, що коефіцієнт використання ресурсів повинен коливатися у межах від 30% до 80%. Оскільки, якщо коефіцієнт використання ресурсів більше 80% починають виникати непередбачувані збої, якщо коефіцієнт використання ресурсів менше 30%, тоді фіксується простій обладнання, та надлишкові витрати на його обслуговування. Тому в процесі моделювання оцінювалася ймовірність того, що ресурси системи використовуються менше заданого порогового значення *a*, а також ймовірність того що ресурси системи використовуються більше заданого порогового значення *b*

$$R_{2j}{}^i \supset R_{2j\_a}{}^i = \{R_{2j}^k | R_{2j}^k < a * V_{R2j}; k = \overline{1, T_{мод}}\} \qquad \left|R_{2j\_a}{}^i\right| = A^i \quad (20)$$

$$R_{2j}{}^i \supset R_{2j\_b}{}^i = \{R_{2j}^k | R_{2j}^k > b * V_{R2j}; k = \overline{1, T_{мод}}\} \qquad \left|R_{2j\_b}{}^i\right| = B^i \quad (21)$$

Таблиця 7.14 – Показники якості та відповідні оцінки якості

Показник якості	Порогове значення	Оцінка	Значення оцінки
$t_з$- час затримки	P_{tj} =0.8сек	p_{1j} $\forall j = \overline{1,M}$	$p_{1j} = 1 - (\sum_{i=1}^{100} k_{ij})/100$ $k_{ij} = \begin{cases} 1 & t_j^i > P_{tj} \\ 0 & \text{інакше} \end{cases}$
	P_{tall}=8 сек	p_{1all}	$p_{1all} = 1 - (\sum_{i=1}^{100} K_{iall})/100$ $K_{iall} = \begin{cases} 1 & t_{all}^i > P_{tall} \\ 0 & \text{інакше} \end{cases}$

$\boldsymbol{P}$ – ймовірність успішного обслуговування	P_{zi}=0,98	p_{2j} $\forall j = \overline{1,M}$	$p_{2j} = 1 - (\sum_{i=1}^{100} K_{iz})/100$ $K_{iz} = \begin{cases} 1 & \frac{N_\Sigma^i - z_j^i}{N_\Sigma^i} > P_{zi} \\ 0 & \text{інакше} \end{cases}$
	P_{zall}=0,98	p_{2all}	$p_{2all} = 1 - (\sum_{i=1}^{100} K_{iall})/100$ $K_{i2all} = \begin{cases} 1 & \frac{N_\Sigma^i - z_{all}^i}{N_\Sigma^i} > P_{zall} \\ 0 & \text{інакше} \end{cases}$
$\boldsymbol{\alpha}$ – коефіцієнт використання ресурсів ГТС	a=0,3	$P_{3R_{2j}}$ $\forall j = \overline{1,M}$	$P_{3R_{2j}} = (\sum_{i=1}^{100} \frac{A_i}{T_{\text{мод}}})/100$
	b=0,8	$P_{4R_{2j}}$ $\forall j = \overline{1,M}$	$P_{4R_{2j}} = (\sum_{i=1}^{100} \frac{B_i}{T_{\text{мод}}})/100$

Таблиця 7.15 Результати моделювання

	Стандартне обслуговування	Керування за методологією	Стандартне обслуговування	Керування за методологією
	Середня затримка обслуговуванні сервісу у вузлі $(\overline{t_j} = \sum_{i=1}^{100} {t_j}^i/100)$		Оцінка вчасного обслуговування у вузлі $(p_1 = (\sum_{j=1}^{M} p_{1j})/M)$	
Max_i	7	9	0,8	0,805
Min_i	1	1		
Середнє	3,8	4,4		
	Середня затримка обслуговуванні сервісу по системі $(\overline{t_{all}} = \sum_i {t_{all}}^i/100)$		Оцінка вчасного обслуговування $(p_{2=}p_{1all})$	
max	80	90	0,82	0,84
Min	20	23		
Середнє	55	62		
	Ймовірність успішного обслуговування сервісу у вузлі $(\overline{z_j} = \sum_i \frac{N_\Sigma^i - z_j^i}{N_\Sigma^i}/100)$		Оцінка ймовірності успішного обслуговування у вузлі $(p_3 = (\sum_{j=1}^{M} p_{2j})/M)$	
Max_i	1	1	0,95	0.99
Min_i	0,96	0.99		
Середнє	0,98	0.999		
	Ймовірність успішного обслуговування сервісу по системі $(\overline{z_{all}} = \sum_i {z_{all}}^i/100)$		Оцінка ймовірності успішного обслуговування по системі $(p_4 = p_{2all})$	

Max_i	1	1	0,94	0.99
Min_i	0,95	0.99		
середнє	0,975	0.999		
	Коефіцієнт використання ресурсів ГТС $\boldsymbol{\alpha}$		Ймовірність використання ресурсу R1 і R2 з коефіцієнтом вик. обч. ресурсів менше заданого порогового значення ($\overline{P_{3R_{1_j}}}$ і $\overline{P_{3R_{2_j}}}$)	
$(Max_j\overline{R_{1j}})/\ V_{R2j}$	0.95	0.9	0.4 і 0,35	0,15 і 0,25
$(Min_j\overline{R_{1j}})/\ V_{R2j}$	0.15	0.25.	Ймовірність використання ресурсу R1 і R2 з коефіцієнтом вик. обч. ресурсів більше заданого порогового значення ($\overline{P_{4R_{1_j}}}$, $\overline{P_{4R_{2_j}}}$)	
$(Max_j\overline{R_{2j}})/\ V_{R2j}$	1	0.85	0,2 і 0,15	0,05 і 0,1
$(Min_j\overline{R_{2j}})/\ V_{R2j}$	0.1	0.2	Оцінка коефіцієнту вик. обч. ресурсів, які використовуються у межах норми $p_5 = 1 - \frac{\sum_{g=1}^{2} w_g p_{3Rgj}}{2} - \frac{\sum_{g=1}^{2} w_g p_{4Rgj}}{2}$	
середнє	0.4	0.2	0.43	0.75

Із таблиці видно, що такі показники якості обслуговування як середня затримка обслуговуванні сервісу у вузлі і в цілому по системі, ймовірність успішного обслуговування сервісу у вузлі та в цілому по системі були утримані у межах допустимих значень та набули незначного покращення. Однак коефіцієнт використання ресурсів при обслуговуванні відповідно до запропонованих моделей та методів утримується у заданих межах із імовірністю більшою в середньому на 32%.

7.6 Імітаційна модель для методу вибору потужності обслуговуючого пристрою

Метод вибору потужності обслуговуючого пристрою для забезпечення процесу обслуговування на заданому рівні якості для багатоканального обслуговування [124].

Метод обслуговування заявок з використанням віртуалізованого обслуговуючого пристрою. Особливість застосування віртуалізованого обслуговуючого пристрою полягає у можливості збільшувати ресурси для виконання обчислювальних операцій. Така можливість з'являється через організацію віртуального обчислювального простору, який не прив'язаний до фізичного обладнання, але залежить від його кластерної організації. На відміну від звичайного підходу щодо балансування навантаження, коли для забезпечення зростаючого навантаження додається додаткові обслуговуючі пристрої між якими виконується розподілення запитів за деяким законом, запропонований метод дозволяє при обслуговуванні телекомунікаційних сервісів, які характеризуються високою інтенсивністю вхідного потоку використовувати набір віртуальних сутностей, робота яких утворює єдиний простір обслуговування, та розрахувати параметри системи як єдиного обслуговуючого пристрою із змінною інтенсивністю обслуговування.

Для розрахунку рекомендованого значення потужності вузла обслуговування, необхідно обрати найбільш відповідні метрики хмарних модулей, підставити значення метрик у модель і провести розрахунок значення потужності вузла обслуговування.

• λ – середнє значення проміжків часу між моментами надходження замовлень,

• n – кількість заявок, які можуть одночасно обслуговуватися у віртуальному вузлі.

• v_k – кількість узагальненого обчислювального ресурсу, який використовується для обслуговування однієї заявки при одночасному обслуговуванні k заявок $k = \overline{1,n}$.

• a, b – верхня та нижня границя завантаженості системи обслуговування,

• R – задане експертами значення, який показує долю часу роботи системи без черги.

V – кількість узагальненого ресурсу

В табл. 7.16 приведено значення, що були взяті на системі обслуговування хмарних обчислень в момент найбільшої завантаженості.

Таблиці 7.16 Вхідні данні моделі

Параметр	Одночасне значение
n	500
λ	648
v_k^g	4Mb
V^g	1612Mb
R	90%
a	0.3
b	0.8

Програмна реалізація моделі

Програмна реалізація виконана за допомогою програмної мови Python. Python – це інтерпретована мова програмування високого рівня для програмування загального призначення. Python має філософію дизайну, яка підкреслює читаність коду, особливо використовуючи значні прогалини. Він надає конструкції, які забезпечують чітке програмування як на малих, так і на великих масштабах.

Файл calc_mu.py призначений для розрахунку імовірностей та визначення цільової функції r.

Лістинг на рис 7.15 відображає програмну реалізацію розрахунку потужності вузла обслуговування. В даній програмі рналізовано обрану математичну модель, значення інтенсивності обробки вхідного потоку заявок розрахованого відповідно до запропонованого методу.

Результат пошуку потужності вузла хмарних обчислень для вказаного прикладу відображено на рис. 7.16.

```python
import numpy as np

N = 500  # кількість заявок, які можуть одночасно обслуговуватись у віртуальному вузлі
V = 1612 # кількість узагальненого ресурсу
VK = 4   # кількість ресурсу, необхідного для обслуговування однієї заявки
A = .3  # нижняя граница завантаженості системи обслуговання
B = .8  # верхня границя завантаженості системи обслуговання

MU_STEP = 0.5 # Точність візначення інтенстивності обслуговування

def calc_speed(k, lmbd, mu):
    return lmbd / (k * mu)


def calc_pa(n, lmbd, mu):
    pa = 1
    for i in range(1, n + 1):
        pa *= calc_speed(i, lmbd, mu)
    return pa


def calc_p0(n, lmbd, mu):
    epa = 0
    for k in range(1, n+1):
        epa += calc_pa(k, lmbd, mu)

    return (1 + epa + (calc_speed(n, lmbd, mu) / (1 - calc_speed(n, lmbd, mu))) * calc_pa(n, lmbd, mu)) ** -1


def calc_pk(k, p0, lmbd, mu):
    return p0 * calc_pa(k, lmbd, mu)


def calc_pnk(n, p0, kk, lmbd, mu):
    return p0 * calc_speed(n, lmbd, mu) ** kk * calc_pa(n, lmbd, mu)


def calc_r(n, mu, lmbd):
    r = 0
    p0 = calc_p0(n, lmbd, mu)
    for k in range(1, n+1):
        r += calc_pk(k, p0, lmbd, mu) / p0

    return r


def calc_v(n, mu, lmbd):
    v = 0
    p0 = calc_p0(n, lmbd, mu)
    for k in range(1, n+1):
        v += VK * k * calc_pk(k, p0, lmbd, mu) / p0

    return v

def cel_f(n, mu, lmbd):
    cel = 0
    p0 = calc_p0(n, lmbd, mu)
    for k in range(1, n+1):
        cel += k*VK*calc_pk(k, p0, lmbd, mu)
    return cel

def main():
    with open('out.txt', 'w') as f:
            l=648 # середнє значення проміжків часу між моментаминадходження замовлень
            for mu in np.arange(l / B, l / A + MU_STEP, MU_STEP):
                rr = calc_r(N, mu, l)
                vv = calc_v(N, mu, l)
                cel=cel_f(N, mu, l)
                res = f"r{l={l}, mu={mu:.2f}, cel={cel}) = {rr}"
                print(res)
                if .89 < rr < .91 and vv <= V:
                    f.write(res + '\n')
                    print('\t!!!')
                    break


if __name__ == '__main__':
    main()
```

Рис. 7.15 Реалізація розрахунку потужності вузла обслуговування

```
r(l=648, mu=989.00, cel=2.6208291203235583) = 0.9255416024328514
r(l=648, mu=989.50, cel=2.6195048004042443) = 0.9249041996794265
r(l=648, mu=990.00, cel=2.6181818181818177) = 0.9242676514439666
r(l=648, mu=990.50, cel=2.6168601716304893) = 0.9236319560439881
r(l=648, mu=991.00, cel=2.6155398587285577) = 0.9229971118013489
r(l=648, mu=991.50, cel=2.614220877458396) = 0.9223631170422353
r(l=648, mu=992.00, cel=2.612903225806452) = 0.9217299700971481
r(l=648, mu=992.50, cel=2.611586901763224) = 0.9210976693008897
r(l=648, mu=993.00, cel=2.6102719033232633) = 0.9204662129925482
r(l=648, mu=993.50, cel=2.6089582284851542) = 0.9198355995154865
r(l=648, mu=994.00, cel=2.6076458752515093) = 0.9192058272173276
r(l=648, mu=994.50, cel=2.6063348416289593) = 0.9185768944499395
r(l=648, mu=995.00, cel=2.6050251256281407) = 0.9179487995694247
r(l=648, mu=995.50, cel=2.603716725263687) = 0.9173215409361055
r(l=648, mu=996.00, cel=2.6024096385542173) = 0.9166951169145093
r(l=648, mu=996.50, cel=2.601103863522328) = 0.9160695258733574
r(l=648, mu=997.00, cel=2.599799398194584) = 0.9154447661855512
r(l=648, mu=997.50, cel=2.5984962406015035) = 0.9148208362281575
r(l=648, mu=998.00, cel=2.597194388777555) = 0.9141977343823974
r(l=648, mu=998.50, cel=2.5958938407611414) = 0.9135754590336321
r(l=648, mu=999.00, cel=2.5945945945945956) = 0.9129540085713501
r(l=648, mu=999.50, cel=2.5932966483241624) = 0.912333381389154
r(l=648, mu=1000.00, cel=2.592000000000001) = 0.9117135758847482
r(l=648, mu=1000.50, cel=2.5907046476761613) = 0.9110945904599248
r(l=648, mu=1001.00, cel=2.589410589410588) = 0.9104764235205524
r(l=648, mu=1001.50, cel=2.5881178232651023) = 0.9098590734765611
    !!!

Process finished with exit code 0
```

Рис. 7.16 Результат розрахунку файлу calc_mu.py

При даних обмеженнях і цільовій функції обчислене значення μ=1001.5 При цьому сама цільова функція має значення 2.58811.

При збільшенні λ значення μ також збільшується, що підтверджує працездатність обраного методу.

Розрахунок оптимального значення інтенсивності вхідного потоку для існуючої системи

Для розрахунку оптимального значения λ проведено аналіз доступних у хмарній системі метрик для найнавантаженішого дня у 2017 році (24.11.2017 «Чорна п'ятниця»). Максимальне значення інтенсивності вхідного потоку було отримано у години найбільшого навантаження, на момент, коли метрики сигналізували про потребу у додатковому ресурсі.

Під час аналізу, для кожного запису була розрахована інтенсивність вхідного потоку (λ) і одночасне значення інтенсивності обслуговування заявки (μ). Базуючись на отриманих даних, можна зробити висновок, що максимальному допустимому значенню інтенсивності вхідного потоку для єдиного процесу, при якому не виникає деградації сервісу (відсутність відмов) відповідає λ = 648 при μ = 1001.5

Дані розрахунку оптимального значення λ і μ занесені в табл. 7.17.

Таблиця 7.17

Дані розрахунку оптимального значення λ и μ

Дата	Час	μ	λ
24.11.2017	00:00	930.5	601.93
	01:00	621.5	401.95
	02:00	587	379.95
	03:00	575	371.96
	04:00	566	365.96
	05:00	603	389.95
	06:00	692.5	447.95
	07:00	850	549.93
	08:00	1375.5	889.89
	09:00	1227.5	793.90
	10:00	1283	829.90
	11:00	1323	855.89
	12:00	1351	873.89
	13:00	1360	879.89
	14:00	1338.5	865.89
	15:00	1331.5	861.89
	16:00	1286	831.90
	17:00	1329	859.89
	18:00	1338,5	865.89
	19:00	1314	849.89
	20:00	1273.5	823.90
	21:00	1131.5	731.91
	22:00	1001.5	647.92
	23:00	859.5	555.93

Графічне відображення розрахованої інтенсивності обслуговування заявки (μ) зображено на рис. 7.17, а графічне відображення одночасного значення інтенсивності обслуговування заявки (μ) і інтенсивності вхідного потоку (λ), для єдиного процесу, зобюражено на рис. 7.18.

Рис. 7.17 Графік значення інтенсивності μ

Розрахунок значення інтенсивності обслуговуваня вхідного потоку заявок для усунення втрат, в системі, де наявні втрати.

Для повноти аналізу проведемо розрахунок інтенсивності обробки потоку вхідних заявок в системі, в якій наявні відмови, для їх усунення.

В табл. 7.18 чітко видно що інтенсивність обробки вхідного потоку заявок, що доступна системі (μ) значно менша за ту, що пропонує використаний метод ($\mu_{роз}$), однак, це обумовлено тим що μ розрахована при умові допустимості черги та відмов, а $\mu_{роз}$ розрахована так щоб гарантовано обслужити всі заявки та мати запас 10% (при R=90%). Для систем хмарних обчислень така така принята доля часу роботи без черги є актуальною, адже при затримках у отриманні провайдером даних з сервісу, спричиненими мережевими затримками, необхідно мати час на прорахунок необхідного значення $\mu_{роз}$ та надсилання керуючого сигналу до моменту початку відмов.

Однак, слід зазначити що $\mu_{роз}$ – найменше можливе значення для обраного методу і при її збільшенні фактична доля часу роботи системи без черги зменшується.

Таблиця 7.18

Данні порівняльного розрахунку μ і $\mu_{роз}$

Дата	Час	Передбачувані відмови при μ	μ	λ	$\mu_{роз}$
11.04.2018	00:00	0	323	645.92	998.5
	01:00	0	217	433.95	671
	02:00	0	205	409.95	634
	03:00	0	200	399.95	618
	04:00	0	198	395.95	612
	05:00	0	211	421.95	652

	06:00	0	241	481.94	744
	07:00	0	296	591.93	915
	08:00	100	478	955.88	1477.5
	09:00	100	427	853.89	1320
	10:00	300	446	891.89	1378.5
	11:00	300	460	919.89	1421
	12:00	400	471	941.88	1456
	13:00	450	474	947.88	1464
	14:00	400	466	931.88	1440.5
	15:00	300	456	911.89	1409
	16:00	300	448	895.89	1385
	17:00	300	449	897.89	1388
	18:00	300	446	891.89	1378.5
	19:00	200	438	875.89	1354
	20:00	0	423	845.90	1307.5
	21:00	0	395	789.90	1221
	22:00	0	350	699.91	1082
	23:00	0	300	599.93	927.5

Графічне відображення розрахованої інтенсивності обслуговування заявкок ($\mu_{роз}$), доступної інтенсивності обслуговування заявкок(μ), відмов при цьому значенні доступної інтенсивності обслуговування заявкок і одночасного значення інтенсивності вхідного потоку (λ), для наявної системи зображено на рис. 7.18.

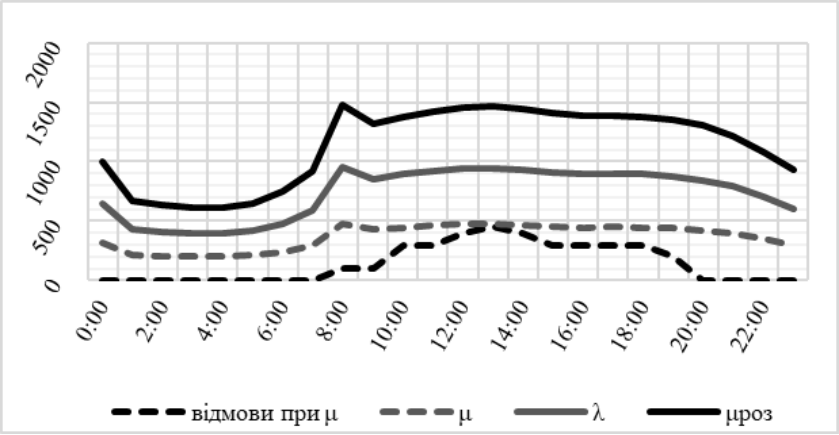


Рис 7.18 Графік $\mu_{роз}$ для запропонованої системи

Розрахунок прогнозу показує показує значне підвищення розрахованої інтенсивності обслуговування заявкок ($\mu_{роз}$) при скачках інтенсивності вхідного потоку заявок(λ), що гарантує роботу системи без черг та відмов.

7.7 Імітаційна модель MATLAB для методу побудови розкладу залучення ресурсів на основі довгострокової статистики із забезпеченням показників якості та енергоефективності

Ітаційна модель методу побудови розкладу залучення ресурсів на основі довгострокової статистики складається з 6 файлів.

1. Файл res.m призначений для розрахунку імовірностей та визначення цільової функції g.

```
function    [g]=res(x)
global l    % порогова кількість запитів у черзі на обслуговування
global m    % максимальна кількість запитів у черзі на обслуговування
global p    % масив імовірностей
global lambda   % значення інтенсивності вхідного потоку
%x(1)=mu - інтенсивність обслуговування заявки

alpha=lambda/4/x(1);

A=1-alpha^(m-l)-(m-l)*(1-alpha)*alpha^(m+2);
p(1)=(1-alpha)*(1-alpha^(m-l))/A;   % значення p0, оскільки нумерація з 1
for i=2:1:(l+2)
   p(i)=alpha^(i-1)*p(1);  % значення pk, 1<=k<=l+1
end
for i=(l+3):1:(m+1)
```

```
    p(i)=(1-alpha)*(alpha^(i-1)-alpha^(m+l))/A; % значення pk, l+2<=k<=m
end
p(m+2)=(1-alpha)^2*alpha^(m+1)/A;
g=x(1);          % цільова функція
```

2. Файл rosenbrock.m призначений для надання всіх .необхідних констант та ініціалізації робочих масивів. В ньому слід задати (записати в текстовому редакторі) наступні константи:

μ^k = (4, 8, 12, 16, 20)– інтенсивність обслуговування заявок за умови k-ї конфігурації обслуговуючого пристрою, що забезпечує показники якості обслуговування на заданому рівні, тобто кількість заявок за одиницю часу, які можуть бути обслужені, k=1,..K, K – кількість можливих конфігурацій системи обслуговування. –

λ - середнє значення проміжків часу між моментами надходження замовлень, незалежні випадкові величини, розподілені за показниковим законом, дані вводяться з файлу 1.xls,

ε=0,05 - мале число

Δ = 0,01c - Допустимий час затримку заявок у процесі обслуговування. –

l = 4 – порогове значення кількості заявок у черзі –

m =5 – максимальна кількість заявок у черзі. –

s1=1, s2=1, s3=2, s4=2, s5=2 для этапа 3

P_{lost}=0,05 – допустимий відсоток втрачених пакетів.

```
function h = rosenbrock(x)
global K    % Кількість конфігурацій
global eps  % похибка
global l    % порогова кількість запитів у черзі на обслуговування
global m    % максимальна кількість запитів у черзі на обслуговування
global Plost    % допустима частка втрачених пакетів
global R    % частка заявок, які обслуговуються у системі не більше доп.часу
global rho  % оптимальне завантеження системи
global s    % кількість значень імовірностей для розрахуку обмежень
global p    % масив імовірностей
% задаємо значення параметрів
K=5;
eps=0.05;
l=4;          % порогова кількість запитів у черзі на обслуговування
m=5;          % максимальна кількість запитів у черзі на обслуговування
p=zeros(1,m+2);
Plost=0.05;     % допустима частка втрачених пакетів
R=0.8;          % відсоток заявок, які обсл.у системі не більше доп.часу
rho=0.8;        % оптимальне завантаження системи
s=[1 1 2 2 2];  % кількість значень імовірностей для розрахуку обмежень
h=res(x);
```

3. Файл unitdisk.m призначений для формування обмежень на кількість ресурсів обслуговування.

```
function [c, ceq] = unitdisk(x)
global K    % Кількість конфігурацій
global mu   % інтенсивність обробки при K-й конфігурації
global Plost % допустима частка втрачених пакетів
global R    % відсоток заявок, які обслуговуються у системі не більше доп.часу
global rho  % оптимальне завантеження системи
global s    % кількість значень імовірностей для розрахуку обмежень
global m    % максимальна кількість запитів у черзі на обслуговування
global p    % масив імовірностей
global lambda   % значення інтенсивності вхідного потоку
%x(1) - номер конфігурації

P=0;
a=floor(x(1));          % ціла частина від номера конфігурації
for i=1:1:s(a)
   P=P+p(i+1);          % сума імовірностей
end
mu=(4*x(1));            %інтенсивність обробки при K-й конфігурації
% Формування вектору неточних обмежень
c(1,1)=lambda/mu-rho;   % обмеження на навантаження системи
c(1,3)=p(m+2)-Plost;    % обмеження на частку втрачених пакетів
c(1,3)=P-R;             % обмеження на частку заявок,обсл.у системі
c(1,4)=x(1)-K;          % обмеження на кількість конфігурацій згори
c(1,2)=-x(1)+1;         % обмеження на кількість конфігурацій знизу
 ceq = [];              % точні обмеження відсутні
```

4. Файл func.m призначений для розрахунку функції f(t).

```
function func()
global data     % масив вхідних даних
global L        % кількість інтервалів усереднення в функції
global DT       % інтервал усереднення
global f        % апроксимуюча функція
data=xlsread('1.xls');  % читаємо файл навчальних даних
n=length(data);
DT=100;                 % інтервал усереднення
K=floor(n/DT);
for k=1:1:DT
   t(k)=k;              % формуємо масив часу
end;
j=1;
for k=1:1:K
   for i=1:1:DT
      x(i)=data(j);   % формуємо масив даних на інтервалі усереднення
```

```
      j=j+1;
   end
   f(k,1)=(k-1)*DT+1;  % початок інтервалу усереднення
   f(k,2)=k*DT;        % кінець інтервалу усереднення
   [a]=polyfit(t,x,1); % лінійна апроксимація МНК
   f(k,3)=a(1);        % нахил
   f(k,4)=a(2);        % зміщення
end
%fprintf('%5.0f %5.0f %3.5f %3.5f\n',f');
% об»єднання інтервалів з близькими показниками апроксимації
eps=0.05;
k=2;
ll=0;
L=K;
while k<L
   if (abs(f(k,3)-f(k-1,3))<=eps) && (abs(f(k,4)-f(k-1,4))<=eps)
      % якщо треба об»єднати інтервали
      f(k-1,2)=f(k,2);            % формуємо попередній рядок f
      f(k-1,3)=(f(k-1,3)+f(k,3))/2;
      f(k-1,4)=(f(k-1,4)+f(k,4))/2;
      L=K-ll;
      for i=(k+1):1:L
         f(i-1,1)=f(i,1);        % підтягуємо f
         f(i-1,2)=f(i,2);
         f(i-1,3)=f(i,3);
         f(i-1,4)=f(i,4);
      end
      for i=(K-ll):1:K
         f(i,1)=0;               % обнуляємо зайве f
         f(i,2)=0;
         f(i,3)=0;
         f(i,4)=0;
      end
       ll=ll+1;
      k=k-1;
   end
   % якщо не треба об»єднати інтервали
   k=k+1;
end
%fprintf('%5.0f %5.0f %3.5f %3.5f\n',f');
```

5. Файл lambda1.m призначений для розрахунку параметрів вхідних даних.

```
function    lambda1(k)
global data     % масив вхідних даних
global L        % кількість інтервалів усереднення в функції
```

```
global lambda   % розраховане значення для інтенсивності вхідного потоку
global sigma    % середньоквадратичне відхилення
global f        % апроксимуюча функція
global N        % кількість перемикань
global nk   % номер інтервалу, на якому треба зробити перемикання
for i=1:1:N
   y(1,i)=i;
   y(2,i)=data(N*(k-1)+i);
end;
sigma=zeros(1,L-1);
for j=1:1:(L-1)
   a=f(j,2)-f(j,1)+1;
   if a<N N=a; N=N; end
   sig=0;
   for i=1:1:N
      sig=sig+(y(2,i)-f(j,3)*i-f(j,4))^2;
   end
   sigma(j)=sqrt(sig/(N-1));
end
nk=find(sigma==min(sigma));
sigma=min(sigma);
lambda=0;
for j=1:1:N
   lambda=lambda+y(2,j);
end
lambda=lambda/N;
```

6. Файл start.m призначений для запуску програми на виконання. Використовується функція оптимізації fmincon.

```
global data     % масив вхідних даних
global sigma    % середньоквадратичне відхилення
global lambda   % розраховане значення для інтенсивності вхідного потоку
global nk       % номер інтервалу, на якому треба зробити перемикання
global f        % апроксимуюча функція
global N        % тривалість інтервалу вимірювання
func            % побудова навчальної функції
%options = optimset('Display','iter','Algorithm','active-set');
n=length(data);
N=100;
K1=floor(n/N);
%K1=5;
swich=zeros(K1,4);
for k=1:1:K1
lambda1(k);      % значення інтенсивності вхідного потоку
[x,fval] = fmincon(@rosenbrock,[1],...
```

```
[],[],[],[],[],[],@unitdisk,options);

k1=floor(x(1)); % конфігурація, на яку треба перейти
mu=4*k1;          % інтенсивність обслуговування в k-й конфігурації
mu0=4*(k1-1);     % інтенсивність обслуговування в k-й конфігурації
N=f(nk,2)-f(nk,1)+1;
%N=20;
for i=1:1:N
   if (mu-(f(nk,3)*i+f(nk,4))*4)<3*sigma break,end;
end
tk=((k-1)*N+i)*5;
swich(k,1)=lambda;
swich(k,2)=sigma;
swich(k,3)=tk;
swich(k,4)=mu;
end
%fprintf('%5.0f %5.0f %3.5f %3.5f\n',f');
fprintf('lambda sigma   tk     mu\n');
fprintf('%5.3f  %5.3f   %5.0f   %5.0f\n',swich');
fprintf('\n');
```

Для запуску програми слід переписати всі вказані файли в робочу папку Matlab і запустити файл start.m.

Результат розрахунку для вказаного прикладу (наведено тількі кінцеві результати):

```
>> start
lambda      sigma      tk      mu
0.870       0.636       35         4
0.320       0.598      1000        4
0.400       0.651      1500        4
0.180       0.405      2000        4
0.070       0.260      7500        4
6.110       2.061      2505        4
0.020       0.141     10500        4
7.050       2.403      3505        8
0.180       0.386      9000        4
4.480       2.338      4505        4
1.770       1.262      5005        4
2.880       1.767      5505        4
5.160       2.243      6005        4
5.340       1.652      6505        4
6.330       1.984      7005        4
6.110       2.061      7505        4
5.860       2.317      8005        4
5.670       2.146      8505        4
```

4.480	2.338	9005	4
7.080	2.095	9505	8
7.700	2.107	10005	8
7.050	2.403	10505	8
5.420	1.621	11005	4
4.260	1.736	11505	4
4.270	1.393	12005	4
4.060	1.376	12505	4
1.970	1.187	13005	4
3.200	1.459	13505	4

Отримано масив моментів переключення та потрібних конфігурацій для кожного λ. Перевірка розрахованого розкладу у імітаційній моделі наведеній у першій частині Додатку 1 за наборами статистичних даних наведених у другій частині Додатку 1 показала, що для кожної статистичної вибірки обчислювальних ресурсів відповідно до імітаційної моделі було достатньо для обслуговування запитів без додаткової затримки.

7.8 Імітаційна модель MATLAB для методу короткострокового прогнозування навантаження

Програма для розрахунку складається з 3 файлів.

1. Файл unitdisk.m призначений для формування обмежень на кількість ресурсів обслуговування.

λ_i - кількість заявок за 1мс, ($i \epsilon 0,..N$), $N=T_{інф}$ /*1мс*, $\lambda_i \epsilon \Lambda$, Λ - множина значень статистики кількості заявок, що надходили протягом часу $T_{інф}$ до початку здійснення прогнозу, $|\Lambda| = N$.

```
function [c, ceq] = unitdisk(x)
global P    % імовірність помилки прогнозування
global p    % розрахована помилка прогнозування
global p2   % розрахована помилка прогнозування
c=p-1+P;    % обмеження на помилку прогнозування
ceq = [p2]; % точні обмеження
```

2. Файл rosenbrock.m призначений для імовірностей та визначення цільової функції g.

```
function [g] = rosenbrock(x)
global data          % масив вхідних даних
global Tn            % інтервал часу, для якого потрібен прогноз
global Tnp           % період прогнозування
global M             % гранична кількість заявок
global P             % імовірність помилки прогнозування
global p             % розрахована помилка прогнозування
global p2            % розрахована помилка прогнозування

Tinf=floor(x(1));        % час збирання інформації
if Tinf>0
DeltaT=Tn(2)-Tn(1);      % тривалість часу, для якого потрібен прогноз
Z=0;                     % лічильник прапорців перемикання
for j=1:(DeltaT-Tinf)
for k=1:Tinf
    t(k)=k;              % формуємо масив часу
    y(k)=data(Tn(1)+j-1+k);  % формуємо масив даних на інтервалі усереднення
end;

[a]=polyfit(t,y,1);      % лінійна апроксимація МНК
sig=0;
for k=1:Tinf
    sig=sig+(y(k)-a(1)*k-a(2))^2;
end
sigma=sqrt(sig/(Tinf-1));
lambda=a(1)*Tnp+a(2)+3*sigma;
if lambda>M
    z=1;
else
    z=0;
end
Z=Z+z;                   % нарощуємо лічильник прапорців перемикання
end
p=Z/(DeltaT-Tinf);  % розрахована помилка прогнозування
g=x(1);                  % цільова функція
%fprintf ('%5.2f  %5.2f\n',Tinf,p-1+P);
else
p2=0;
g=0;
end
```

3. Файл start.m призначений для запуску програми на виконання та надання всіх необхідних констант Використовується функція оптимізації fmincon.

В ньому слід задати (записати в текстовому редакторі) наступні константи:

- M=2 – гранична кількість заявок, яка може бути обслужена при заданій конфігурації обслуговуючого пристрою.

- P=0,1 – імовірність помилки прогнозування.

- $T_{пр}$ = 10мс- період прогнозування, час, по завершенню якого запускається алгоритм прогнозування.

Слід відзначити, що алгоритм пошуку мінімального значення $T_{інф}$ дуже залежить від початкових значень для ітерації. Тому доцільно спочатку зробити грубу оцінку $T_{інф}$,

використовуючи перебір можливих значень $T_{інф}$ (закоментований фрагмент), а потім задати максимальне значення з обраного діапазону та запустити програму fmincon.

```
global data % масив вхідних даних
global Tn   % інтервал часу, для якого потрібен прогноз
global Tnp  % період прогнозування
global M    % гранична кількість заявок
global P    % імовірність помилки прогнозування
global p2   % розрахована помилка прогнозування
% задаємо значення параметрів
M=2;        % гранична кількість заявок
P=0.1;      % імовірність помилки прогнозування
Tn=[500 2000];      % початкове значення вхідної послідовності
Tnp=10;     % період прогнозування
Z=0;        % лічильник прапорців перемикання
p2=1;

options = optimset('Display','iter','Algorithm','interior-point');
data=xlsread('1.xls');  % читаємо файл навчальних даних
n=length(data);

[x,fval] = fmincon(@rosenbrock,[570],...
                   [],[],[],[],[],[],@unitdisk,options);
%for i=100:50:1000
%rosenbrock(i);
%end
x(1)
```

Результат виконання програмної реалізації вирішення поставленої задачі визначив оптимальний $T_{інф}$ для досліджуваних даних.

```
Iter F-count            f(x)  Feasibility   optimality         step
    0       2   5.700000e+002   1.000e+000   1.000e+000
    1       4   5.692224e+002   1.000e+000   1.000e+000   7.776e-001
    2      10   5.667463e+002   1.000e+000   1.000e+000   2.476e+000
    3      13   5.666173e+002   1.000e+000   1.000e+000   1.289e-001
    4      15   5.659660e+002   1.000e+000   1.000e+000   6.513e-001
    5      21   5.636402e+002   1.000e+000   1.000e+000   2.326e+000
    6      27   5.469985e+002   1.000e+000   1.000e+000   1.664e+001
    7      30   5.468728e+002   1.000e+000   1.000e+000   1.258e-001
    8      32   5.459751e+002   1.000e+000   1.000e+000   8.976e-001
    9      34   5.388083e+002   1.000e+000   1.000e+000   7.167e+000

Local minimum possible. Constraints satisfied.

ans =

5.388083e+002
```

Таким чином, значення оптимального $T_{інф}$, при якому для даного прикладу виконуються всі умови, складає 538.

Результат роботи запропонованого методу з оптимальним $T_{інф}$

В експерименті було вибрано значення $T_{інф}$=538. Звернувши увагу на прогнозовані дані (рис. 7.19) та реальні (рис. 7.20), можна виділити наступний висновок: результат роботи запропонованого алгоритму відповідає реальним даним.

Після більш детальної оцінки результатів було виявлено, що:

1. 80% відсотків спрогнозованих даних виявились вірними.
2. В 95% відсотках запропонований метод вдало запропонував змінити конфігурацію обладнання.
3. Час виконання алгоритму склав 25 секунд.

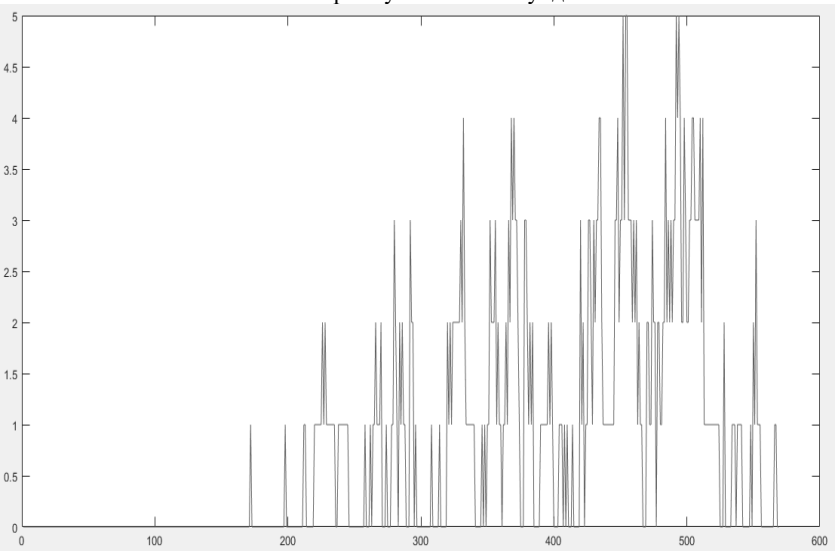


Рис 7.19 Результат роботи запропонованого методу

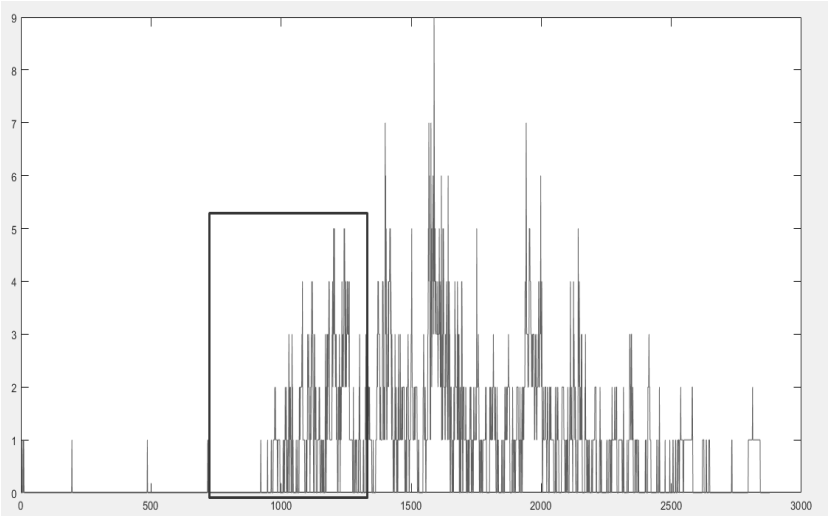


Рис 7.20 Реальні данні

В наступному експерименті значення $T_{інф}$ було фіксованим та вибрано значення 750. Після проведення аналогічних кроків було отримано наступні результати.

1.74% відсотків спрогнозованих даних виявились вірними.

2.В 89% відсотках метод вдало запропонував змінити конфігурацію обладнання.

3.Час виконання алгоритму склав 37 секунд.

Результати показали, що метод з фіксованим $T_{інф}$ також має високі показники правдоподібності прогнозування та вчасної зміни конфігурації обладнання. Але час виконання програмної частини значно вищий від запропонованої варіації. Так як час саме в короткостроковому прогнозування є важливим критерієм, то метод з оптимальним $T_{інф}$ значно кращий для вчасної реакції на сплески навантаження.

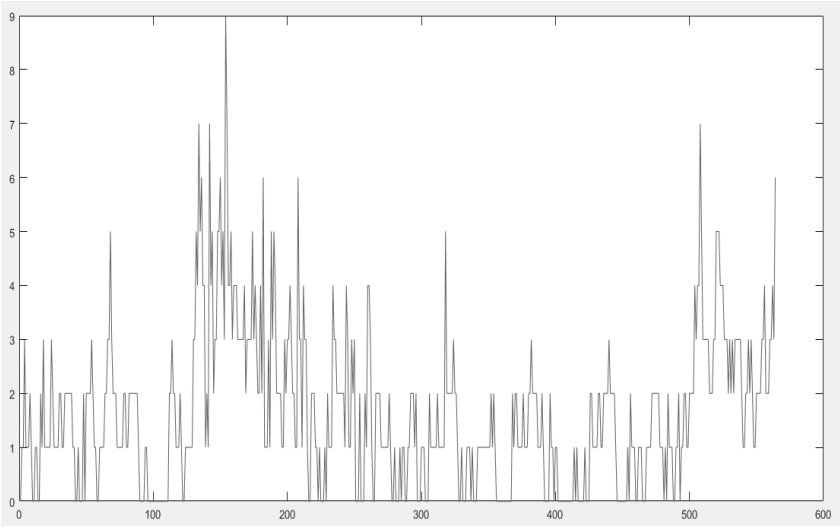


Рис. 7.21 Результат роботи методу з фіксованим $T_{інф}$

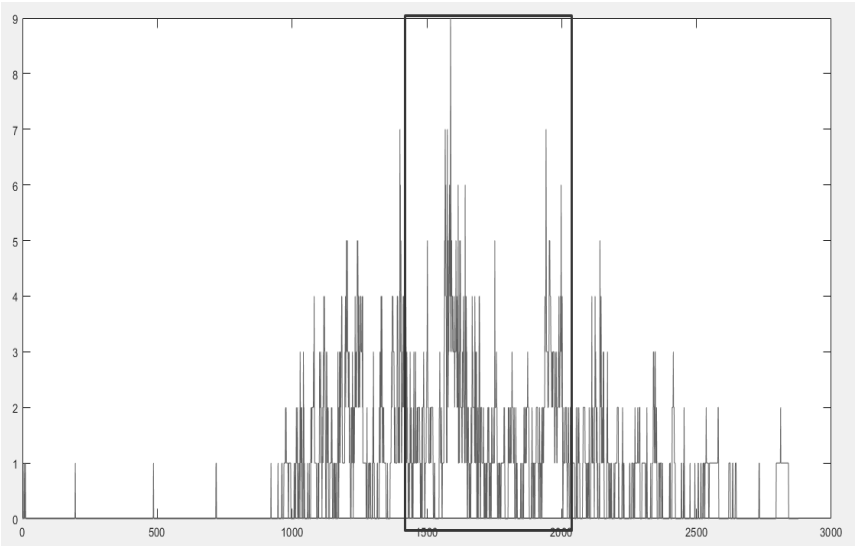


Рис 7.22 Реальні данні

ЗАГАЛЬНІ ВИСНОВКИ

1. Наведено аналіз поточної ситуації на ринку безпроводового зв'язку показує збільшення службового навантаження, що призводить до збільшення необхідності у додаткових ресурсах, разом з тим нерівномірність завантаження вузлів інфраструктури призводить до їх простою, таким чином виникає потреба у впровадженні технологій, які як не призводять до простоїв обладнання, так і гарантують якість обслуговування навантаження протягом дня.
2. Проведено огляд технології віртуалізації NFV показав, що вона є доцільною для побудови безпроводових мереж майбутнього, оскільки забезпечує необхідну гнучкість та масштабованість.
3. Представлено метод визначення місця розміщення та обсягів зарезервованих обчислювальних ресурсів віртуальних мережевих функцій у дата центрах оператора мобільного зв'язку, який гарантує якість надання телекомунікаційних сервісів з мінімально необхідними затратами ресурсів, виконуючи визначення їх достатньої конфігурації, що дозволяє скоротити витрати на 13% у порівнянні з випадково обраною монохмарою та на 47% у порівнянні з традиційним підходом до розгортання мережі.
4. Представлено метод визначення розміру інтервалу часу сталої конфігурації обчислювальних ресурсів, який передбачає його змінну величину і забезпечує гнучке використання ресурсів у віртуалізованому середовищі, зменшуючи відсоток вільних ресурсів на 42% у порівнянні з виділеним обладнанням і на 9% у порівнянні з існуючими аналогами та зменшуючи робоче навантаження у мережі.
5. Представлено розподілений метод локальної реконфігурації обчислювальних ресурсів віртуальної мережі у випадку відмови або перевантаження, який за рахунок децентралізованого керування та врахування міграційних витрат перерозподіляє віртуальні мережеві функції в нормальному та аварійному режимі із забезпеченням економічно обґрунтованого використання ресурсів, зменшуючи витрати в середньому на 21%.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1.Рекомендации ITU-T M.3371 (10/2016) Требования к управлению услугами в системе управления электросвязью, совместимой с облаком. [Електронний ресурс]. – Режим доступу: https://www.itu.int/ITU-T/recommendations/rec.aspx?rec=13064&lang=ru.

2.Y.3500 : Information technology - Cloud computing - Overview and vocabulary. [Електронний ресурс]. – Режим доступу: https://www.itu.int/rec/T-REC-Y.3500-201408-I/en.

3.ETSI GS NFV 003: "Network Functions Virtualisation (NFV); Terminology for Main Concepts in. NFV".

4.Cloud computing framework for end to end resource management Recommendation ITU-T Y.3520. [Електронний ресурс]. – Режим доступу: https://www.itu.int/rec/T-REC-Y.3520/en.

5.ITU-T Y.2011 : General principles and general reference model for Next Generation Networks. [Електронний ресурс]. – Режим доступу: https://www.itu.int/rec/T-REC-Y.2011-200410-I/en.

6.ITU-T Y.3511 (03/14) : Framework of inter-cloud computing. [Електронний ресурс]. – Режим доступу: https://www.itu.int/rec/T-REC-Y.3511/en.

7.Bradai A. Cellular software defined networking: a framework / A. Bradai, K. Singh, T. Ahmed, T. Rasheed // IEEE Communications Magazine. – 2015. – Vol. 53, No. 6. – P. 36-43.

8.Basta A. Applying NFV and SDN to LTE mobile core gateways, the functions placement problem / A. Basta, W. Kellerer, M. Hoffmann, H. Morper et al. // 4th workshop on All things cellular: operations, applications, & challenges. – Chicago, USA, 2014.– P. 33-38.

9.Yousaf F. Z. SoftEPC – Dynamic instantiation of mobile core network entities for efficient resource utilization / F. Z. Yousaf, J. Lessmann, P. Loureiro, S. Schmid // 2013 IEEE International Conference on Communications (ICC). – Budapest, Hungary, 2013.– P. 3602-3606.

10. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2017–2022 White Paper Cisco public. https://s3.amazonaws.com/media.mediapost.com/uploads/CiscoForecast.pdf (2019).

11. Kopetz H. Internet of Things. In: Real-Time Systems. – Springer, Boston, MA, 2011. – P. 307-323.

12. 5G White paper [Online]. – Available at: https://www.nttdocomo.co.jp/english/binary/pdf/corporate/technology/whitepaper_5g/DOCOMO_5G_White_Paper.pdf

13. The Internet of everything – A $19 trillion opportunity [Online]. – Available at: http://www.cisco.com/web/services/portfolio/consulting-services/documents/consulting-services-capturing-ioe-valueaag.pdf

14. Emmerson B. M2M: The internet of 50 billion devices / B. Emmerson // WinWin Magazine. – 2010. – P. 19-22.

15. Shimojo T. Future mobile core network for efficient service operation / T. Shimojo, Y. Takano, A. Khan, S. Kaptchouang, M. Tamura, S. Iwashina // 1st IEEE Conference on Network Softwarization (NetSoft). – London, UK, 2015.– P. 1-6.

16. Shimojo T. Cost-efficient method for managing network slices in a multi-service 5G core network / T. Shimojo, M. R. Sama, A. Khan, S. Iwashina // Proceedings of the 2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM). – Lisbon, Portugal, 2017. – P. 1121-1126.

17. Sama M. R. Reshaping the mobile core network via function decomposition and network slicing for the 5G Era / M. R. Sama, X. An, Q. Wei, S. Beker // Proceedings of the 2016 IEEE Wireless Communications and Networking Conference Workshops (WCNCW). – Doha, Qatar, 2016. – P. 90-96.

18. Mijumbi R. Network Function Virtualization: State-of-the-art and Research Challenges / R. Mijumbi, J. Serrat, J. Gorricho, N. Bouten et al. // IEEE Communications Surveys & Tutorials. – 2015. – Vol. 18, No. 1. – P. 236-262.

19. Hawilo H. NFV: state of the art, challenges, and imple-mentation in next generation mobile networks (vEPC) / H. Hawilo, A. Shami, M. Mirahmadi, R. Asal // IEEE Network. – 2014. – Vol. 28, No. 6. – P. 18-26.

20. Liquid Core [Online]. – Available at: http://networks.nokia.com/portfolio/liquidnet/liquidcore.

21. Soares J. Cloud4nfv: A platform for virtual network functions / J. Soares, M. Dias, J. Carapinha, B. Parreira, S. Sargento // 2014 IEEE 3rd International Conference on Cloud Networking (CloudNet). – Luxembourg, 2014. – P. 288-293.

22. Moens H. VNF-P: A model for efficient placement of virtualized network functions / H. Moens, F. De Turck // 10th International Conference on Network and Service Management. – Rio de Janeiro, 2014. – P. 418-423.

23. Future Communication Architecture For Mobile Cloud Services. Overall Architecture Definition [Online]. – Available at: http://www.mobile-cloud-networking.eu/site/index.php?process=download&id=124&code=93b79f8f5b99f67a6cdc28369c05b65f624cfee7

24. ETSI Industry Specification Group (ISG), “Draft Documents.” Aug2015 [Online]. – Available at: https://docbox.etsi.org/isg/nfv/open/Drafts/

25. Signaling is growing 50% faster than data traffic [Online]. – Available at: http://docplayer.net/6278117-Signaling-is-growing-50-faster-than-data-traffic.html

26. Mueller J. Mobile Cloud – Combining EPC, SDN and NFV / J. Mueller, T. Magendaz // Mobile Network (Function) Virtualization and Software Defined Networks of ITG Informationstechnische Gesellschaft im VDE FG 5.2.4 'Mobility in IP-based Networks'. – Munich, Germany, 2013.

27. Mijumbi R. On the energy efficiency prospects of network function virtualization / R. Mijumbi, J. Serrat, J. Gorricho and J. Rubio-Loyola [Online]. – Available at: https://arxiv.org/pdf/1512.00215.pdf

28. Network Functions Virtualisation (NFV); Use Cases [Online]. – Available at: http://www.etsi.org/deliver/etsi_gs/nfv/001_099/001/01.01.01_60/gs_nfv001v010101p.pdf.

29. The Rise of Virtual EPC: A Mobile Packet Core Forecast & Analysis [Online]. – Available at: http://www.heavyreading.com/details.asp?sku_id=3187&skuitem_itemid=1558

30. Economic Benefits of Virtualized Evolved Packet Core [Online]. – Available at: https://www.cisco.com/c/dam/en/us/solutions/collateral/service-provider/ultra-services-platform/idc-us41419916.pdf

31. The LTE Network Architecture [Online]. – Available at: http://www.cse.unt.edu/~rdantu/FALL_2013_WIRELESS_NETWORKS/LTE_Alcatel_White_Paper.pdf

32. 3GPP Technical Specification 23.203, Policy and charging control architecture [Online]. – Available at: http://www.etsi.org/deliver/etsi_ts/123200_123299/123203/12.06.00_60/ts_123203v120600p.pdf

33. Sesia S. LTE – The UMTS Long Term Evolution: From Theory to Practice / S. Sesia, I. Toufik, M. Baker. – Wiley, 2009. – 792 p.

34. Herrera J. G. Resource Allocation in NFV: A Comprehensive Survey / J. G. Herrera, J. F. Botero // IEEE Transactions on Network and Service Management. – 2016. – Vol. 13, No. 3. – P. 518-532.

35. Network Functions Virtualisation (NFV); Architectural Framework [Online]. – Available at: http://www.etsi.org/deliver/etsi_gs/nfv/001_099/002/01.01.01_60/gs_nfv002v010101p.pdf.

36. ETSI NFV Management & Orchestration (MANO) простым языком [Електронний ресурс] : [Веб-сайт]. – Електронні дані. – Режим доступа: https://sdnblog.ru/etsi-nfv-mano-beginners-tutorial/]. – Назва з екрана.

37. Скулиш М. А. Метод складання розкладу залучення ресурсів для високонавантажених інформаційних систем / М. А. Скулиш // Наукові записки Українського науково-дослідного інститут зв'язку. – 2014. – № 6. – С. 65-70.

38. Sitaram D. Moving To The Cloud: Developing Apps in the New World of Cloud Computing / D. Sitaram, G. Manjunath. – Massachusetts: Syngress, 2011. – 468 p.

39. Vu H.T. A Traffic and Power-aware Algorithm for Virtual Machine Placement in Cloud Data Center / H.T. Vu, S. Hwang // International Journal of Grid and Distributed Computing. – 2014. – Vol. 7, No. 1. – pp. 21-32.

40. M. Peter and G. Timothy. The NIST Definition of Cloud Computing, Recommendations of the National Institute of Standards and Technology [Online]. – Available at: http://www.nist.gov/itl/cloud/.

41. Scharf M. Monitoring and abstraction for networked clouds / M. Scharf, T. Voith, W. Roome, B. Gaglianello, M. Steiner, V. Hilt, and V. Gurbani // 2012 16th International Conference on Intelligence in Next Generation Networks (ICIN). – Berlin, Germany, 2012.– P. 80-85.

42. Mandal A. Provisioning and Evaluating Multi-domain Networked Clouds for Hadoop-based Applications / A. Mandal, Y. Xin, I. Baldine, P. Ruth, C. Heerman, J. Chase, V. Orlikowski, and A. Yumerefendi // 2011 IEEE Third International Conference on Cloud Computing Technology and Science (CloudCom). – Athens, Greece, 2011.– P. 690-697.

43. Lopez D. Network functions virtualization: Beyond carrier-grade clouds / D. Lopez // Optical Fiber Communications Conference and Exhibition (OFC). – San Francisco, CA, USA, 2014.– P. 1-18.

44. H. Basilier M. Darula and J. Wilke. Virtualizing network services the telecom cloud. Techinical White Paper [Online]. – Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.488.5158&rep=rep1&type=pdf

45. Distributed Management Task Force, “Cloud Management Working Group (CMWG),” [Online]. – Available at: http://www.dmtf.org/standards/cmwg

46. McKeown N. OpenFlow: Enabling Innovation in Campus Networks / N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner // SIGCOMM Comput. Commun. Rev. – 2008. – Vol. 38, No. 2. – pp. 69-74.

47. Mauro D.R. Essential SNMP / D. R. Mauro and K. J. Schmidt. – O'Reilly Media, 2005. – 464 p.

48. Wedge G. Carrier-Grade: Five Nines, the Myth and the Reality / G. Wedge and L. Barbara // Pipeline Publications. – 2006. – Vol. 3, No. 1. – P. 1-14.

49. Vishwanath K. V. Characterizing Cloud Computing Hardware Reliability / K. V. Vishwanath and N. Nagappan // ACM Symposium on Cloud Computing. – New York, NY, USA, 2010.– P. 193-204.

50. Network Equipment-Building System (NEBS) [Online]. – Available at: http://en.wikipedia.org/wiki/Network Equipment-Building System

51. Adamuz-Hinojosa O. Automated Network Service Scaling in NFV: Concepts, Mechanisms and Scaling Workflow / O. Adamuz-Hinojosa, J. Ordonez-Lucena, P. Ameigeiras, J. J. Ramos-Munoz, D. Lopez, and J. Folgueira // IEEE Communications Magazine. – 2018. – Vol. 56, No. 7. – P. 162-169.

52. Network Functions Virtualisation (NFV); Management and Orchestration [Online]. – Available at: https://www.etsi.org/deliver/etsi_gs/NFV-MAN/001_099/001/01.01.01_60/gs_NFV-MAN001v010101p.pdf.

53. Network Functions Virtualisation (NFV); Virtual Network Functions Architecture [Online]. – Available at: https://www.etsi.org/deliver/etsi_gs/NFV-SWA/001_099/001/01.01.01_60/gs_NFV-SWA001v010101p.pdf.

54. Network Functions Virtualisation (NFV) Release 2; Management and Orchestration; Network Service Templates Specification [Online]. – Available at: https://www.etsi.org/deliver/etsi_gs/NFV-IFA/001_099/014/02.04.01_60/gs_nfv-ifa014v020401p.pdf.

55. Cheng X. Virtual network embedding through topology-aware node ranking / X. Cheng, S. Su, Z. Zhang, H. Wang, F. Yang, Y. Luo, J. Wang // ACM SIGCOMM Computer Communication Review. – 2011. – Vol. 41, No. 2. – P. 38-47.

56. Zhang S. Virtual network embedding with substrate support for parallelization / S. Zhang, J. Wu, and S. Lu // IEEE GLOBECOM. – Anaheim, CA, USA, 2012.– P. 2615-2620.

57. Gandhi A. Minimizing Data Center SLA Violations and Power Consumption via Hybrid Resource Provisioning / A. Gandhi, Y. Chen, D. Gmach, M. Arlitt, and M. Marwah // Green Computing : International Conference and Workshops, Orlando, FL, 25–28 July 2011 : proceedings. – IEEE, 2011. – P. 1–8.

58. Xu Z. Efficient virtual network embedding via exploring periodic resource demands / Z. Xu, W. Liang, Q. Xia // 2014 IEEE 39th Conference on Local Computer Networks (LCN). – Edmonton, AB, Canada, 2014.– P. 90-98.

59. Beloglazov A. Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in Cloud data centers / A. Beloglazov, R. Buyya // Concurrency and Computation: Practice & Experience. – 2012. – Vol. 24, No. 13. – P. 1397-1420.

60. Gandhi A. Server farms with setup costs / A. Gandhi, M. Harchol-Balter, I. Adan // Performance Evaluation. – 2010. – Vol. 67. – P. 1123-1138.

61. Coskun A. Evaluating the impact of job scheduling and power management on processor lifetime for chip multiprocessors / A. Coskun, R. Strong, D. Tullsen, S. Rosing // ACM SIGMETRICS. – WA, USA, 2009. – P. 169-180.

62. Khuller S. Energy efficient scheduling via partial shutdown / S. Khuller , J. Li , B. Saha // Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms. – Texas, USA, 2010. – pp. 1360-1372.

63. Rajan A.S. Understanding the bottlenecks in virtualiz-ing cellular core network functions / A.S. Rajan, S. Gobriel, C. Maciocco, K.B. Ramia et al. // 2015 IEEE International Workshop on Local and Metropolitan Area Networks (LANMAN). – Beijing, 2015.– P. 1-6.

64. Ferrer Riera J. Virtual network function scheduling: Concept and challenges / J. Ferrer Riera, E. Escalona, J. Ba-talle, E. Grasa, J. A. Garcia-Espin // 2014 International Conference on Smart Communications in Network Technologies (SaCoNeT). – Vilanova i la Geltru, 2014. – pp. 1-5.

65. Jennings B. Resource management in clouds: Survey and research challenges / B Jennings, R Stadler // Journal of Network and Systems Management. – 2014. – P. 1-53.

66. Fischer A. Virtual Network Embedding: A Survey / A. Fischer, J. Botero, M. Till Beck, H. de Meer and X. Hesselbach // IEEE Communications Surveys & Tutorials. – 2013. – Vol. 15, No. 4. – P. 1888-1906.

67. Baumgartner A. Mobile core network virtualization: A model for combined virtual core network function placement and topology optimization / A. Baumgartner, V.S. Reddy, T. Bauschert // 2015 1st IEEE Conference on Network Softwari-zation (NetSoft). – London, 2015. – P. 1-9.

68. Mehraghdam S. Specifying and placing chains of virtual network functions / S. Mehraghdam, M. Keller, H. Karl // 2014 IEEE 3rd International Conference on Cloud Network-ing (CloudNet). – Luxembourg, 2014. – P. 7-13.

69. Lischka J. A virtual network mapping algorithm based on subgraph isomorphism detection / J. Lischka, H. Karl // Proceedings of the 1st ACM workshop on Virtualized infra-structure systems and architectures. – 2009. – P. 81-88.

70. Heller B. The controller placement problem / B. Heller, R. Sherwood, N. McKeown // ACM HotSDN. – Helsinki, Fin-land, 2012. – P. 1-6.

71. Hock D. Pareto-optimal resilient controller placement in SDN-based core networks / D. Hock, M. Hartmann, S. Gebert, M. Jarschel et al. // 25th International Teletraffic Congress (ITC). – Shanghai, 2013. – P. 1-9.

72. Lange S. Heuristic Approaches to the Controller Placement Problem in Large Scale SDN Networks / S. Lange, S. Gebert, T. Zinner, P. Tran-Gia et al. // IEEE Transactions on Network and Service Management. – 2015. – Vol. 12, No. 1. – P. 4-17.

73. Chandra A. Dynamic resource allocation for shared data centers using online measurements / A. Chandra, W. Gong, P. Shenoy // 11th international conference on Quality of service. – Berkeley, CA, USA, 2003. – P. 381-398.

74. Infuhr J. Introducing the virtual network mapping problem with delay, routing and location constraints / J. Infuhr, and G. R. Raidl // 5th international conference on Network optimization. – Springer-Verlag, 2011. – P. 105-117.

75. Jarray A. VCG auction-based approach for efficient Virtual Network embedding / A. Jarray and A. Karmouch // IFIP/IEEE International Symposium on Integrated Network Management. – Ghent, Belgium, 2013. – P. 609-615.

76. Cai Z. Virtual network embedding for evolving networks / Z. Cai, F. Liu, N. Xiao, Q. Liu, and Z. Wang // IEEE Global Telecommunications Conference. – Miami, FL, USA, 2010. – P. 1-5.

77. Sun G. A cost efficient framework and algorithm for embedding dynamic virtual network requests / G. Sun, H. Yu, V. Anand, L. Li // Future Generation Computer Systems. – 2013. – Vol. 29, No. 5. – P. 1265-1277.

78. Mijumbi R. Design and Evaluation of Learning Algorithms for Dynamic Resource Management in Virtual Networks / R. Mijumbi, J.-L. Gorricho, J. Serrat, M. Claeysy, F. D. Turcky, S. Latr // Network Operations and Management Symposium (NOMS), Krakow, 5–9 May 2014. – IEEE, 2014. – P. 1–9.

79. Patikirikorala T. A multi-model framework to implement self-managing control systems for QoS management / T. Patikirikorala, A. Colman, J. Han, and L. Wang // 6th International Symposium on Software Engineering for Adaptive and Self-Managing Systems. – Waikiki, Honolulu, HI, USA, 2011. – P. 218-227.

80. Lai W. Game Theoretic Distributed Dynamic Resource Allocation with Interference Avoidance in Cognitive Femtocell Networks / W. Lai, M. Chiang, S. Lee, T. Lee // IEEE Wireless Communications and Networking Conference. – Shanghai, China, 2013. – P. 3364-3369.

81. Jokhio F. Prediction-Based Dynamic Resource Allocation for Video Transcoding in Cloud Computing / F. Jokhio, A. Ashraf, S. Lafond, I. Porres, J. Lilius // 21st Euromicro International Conference on Parallel, Distributed, and Network-Based Processing. – Belfast, UK, 2012. – P. 254-261.

82. Castellanos M. iBOM: A Platform for Intelligent Business Operation Management / M. Castellanos, F. Casati, M. Shan, U. Dayal // Int. Conf. on Data Engineering (ICDE). – Tokyo, Japan, 2005. – P. 1084-1095.

83. Xu J. Autonomic resource management in virtualized data centers using fuzzy logic based approaches / J. Xu, M. Zhao, J. Fortes, R. Carpenter, M. Yousif // Cluster Computing Journal. – 2008. – Vol. 11, No. 3. – P.213-227.

84. Urgaonkar B. Agile dynamic provisioning of multi-tier Internet applications / B. Urgaonkar, P. Shenoy, A. Chandra, P. Goyal, T. Wood // ACM Transactions on Autonomous and Adaptive Systems (TAAS). – 2008. – Vol. 3, No. 1. – P.1-39.

85. Коваль Б. В. Дослідження площини управління програмно-керованих мереж на основі розподіленої системи функцій віртуалізації / Б. В. Коваль, М. О. Селюченко, Г. В. Мельник, А. В. Ковальчук // Вісник Національного університету «Львівська політехніка». Радіоелектроніка та телекомунікації. - 2014. - № 796. - С. 164-175.

86. Xiong G. A virtual service placement approach based on improved quantum genetic algorithm / G. Xiong, Y.-X. Hu, L. Tian, J.-L. Lan, J.-F. Li, And Q. Zhou // Frontiers of Information Technology & Electronic Engineering. – 2016. – Vol. 17, No. 7. – P. 661–671.

87. Gember-Jacobson A. OpenNF: Enabling Innovation in Network Function Control / A. Gember-Jacobson, R. Viswanathan, C. Prakash, R. Grandl, J. Khalid, S. Das, and

A. Akella // Proceedings of the 2014 ACM Conference on SIGCOMM, ser. SIGCOMM '14. – New York, NY, USA, 2014. – P. 163-174.

88. Rajagopalan S. Split/Merge: System Support for Elastic Execution in Virtual Middleboxes / S. Rajagopalan, D. Williams, H. Jamjoom, and A. Wareld // 10th USENIX Symposium on Networked Systems Design and Implementation (NSDI 13). – Lombard, IL, USA, 2013. – P. 227-240.

89. Xia J. Optimized Virtual Network Functions Migration for NFV/ J. Xia, Z. Cai, M. Xu // IEEE 22nd International Conference on Parallel and Distributed Systems. – Wuhan, China, 2016. – P. 340-346.

90. МСЭ-Т E.800 [Электронный ресурс]. – Режим доступа: https://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-E.800-200809-I!!PDF-R&type=items.

91. Scholler M. Resilient deployment of virtual network functions / M. Scholler, M. Stiemerling, A. Ripke, R. Bless // Proc. 5th Int. Congr. Ultra Mod. Telecommun. Control Syst. Workshops (ICUMT). – St. Petersburg, Russia, 2013. – P. 208-214.

92. Taleb T. On Service Resilience in Cloud-Native 5G Mobile Systems / T. Taleb, A. Ksentini, B. Sericola // IEEE Journal on Selected Areas in Communications. – 2016. – Vol. 34, No. 3. – P. 483-496.

93. Chowdhury N.M.M.K. Virtual Network Embedding with Coordinated Node and Link Mapping / N.M.M.K. Chowdhury, M. Rahman and R. Boutaba // INFCOM. – Rio de Janeiro, Brazil, 2009. – P. 1-9.

94. Fajjari I. VNR Algorithm: A Greedy Approach For Virtual Networks Reconfigurations / I. Fajjari, N. Aitsaadi, G. Pujolle and H. Zimmermann. // IEEE Globel Communications Conference, Exhibition and Industry Forum. – Houston, United States, 2011. – P. 1-7.

95. Rahman M.R. Survivable Virtual Network Embedding / M. R. Rahman, I. Aib, and R. Boutaba // NETWORKING 2010. – 2010. – P. 40-52.

96. Abid H. A novel scheme for node failure recovery in virtualized networks / H. Abid; N. Samaan // 2013 IFIP/IEEE International Symposium on Integrated Network Management (IM 2013). – Ghent, Belgium, 2013. – P. 1154-1160.

97. Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); General Packet Radio Service (GPRS); Service description; Stage 2 [Online]. – Available at: http://www.etsi.org/deliver/etsi_ts/123000_123099/123060/10.03.00_60/ts_123060v100300p.pdf

98. LTE; General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access [Online]. – Available at: http://www.etsi.org/deliver/etsi_ts/123400_123499/123401/12.06.00_60/ts_123401v120600p.pdf

99. Universal Mobile Telecommunications System (UMTS); LTE; Architecture enhancements for non-3GPP accesses [Online]. – Available at: http://www.etsi.org/deliver/etsi_ts/123400_123499/123402/10.04.00_60/ts_123402v100400p.pdf

100. Olsson M. EPC and 4G Packet Networks / M. Olsson, C. Mulligan. – Academic Press, 2012. – 624 p.

101. Prados-Garzon J. Modeling and Dimensioning of a Virtualized MME for 5G Mobile Networks / J. Prados-Garzon, J. J. Ramos-Munoz, P. Ameigeiras, P. Andres-Maldonado, J. M. Lopez-Soler // IEEE Transactions on Vehicular Technology. – 2017. – Vol. 66, No. 5. – P. 4383-4395.

102. Mijumbi R. Server placement and assignment in virtualized radio access networks / R. Mijumbi, J. Serrat, J.-L. Gorricho, J. Rubio-Loyola, S. Davy // 2015 11th International Conference on Network and Service Management (CNSM). – Barcelona, 2015. – P. 398-401.

103. Cau E. Efficient Exploitation of Mobile Edge Computing for Virtualized 5G in EPC Architectures / E. Cau, M. Corici, P. Bellavista, L. Foschini, G. Carella, A. Edmonds, T. M. Bohnert // 2016 4th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering. – Oxford, UK, 2016. – P. 100-109.

104. Network time protocol version 4: Protocol and algorithms specification [Online]. – Available at: https://tools.ietf.org/html/rfc5905

105. Jeon S. Virtualised EPC for on-demand mobile traffic offloading in 5G environments / S. Jeon, D. Corujo, R. L. Aguiar // 2015 IEEE Conference on Standards for Communications and Networking (CSCN). – Tokyo, Japan, 2016. – P. 275-281.

106. Фелижанко А. Виртуализация в сетях мобильной связи [Электронный ресурс]. – Режим доступа: https://www.slideshare.net/CiscoRu/celc-virtualization-inmobilenetworks17042014afelizha

107. Bianco A. Cost and performance trade-offs in reconfiguration strategies for WDM networks / A. Bianco, J. Finochietto, C. Piglione// 4th International Telecommunication Networking Workshop on QoS in Multiservice IP Networks. – Venice, Italy, 2008. – P. 95-100.

108. Chen Y. Managing Server Energy and Operational Costs in Hosting Centers / Y. Chen, A. Das, W. Qin, A. Sivasubramaniam, Q. Wang, N. Gautam // SIGMETRICS '05 Proceedings of the 2005 ACM SIGMETRICS international conference on Measurement and modeling of computer systems. – Banff, Alberta, Canada, 2005. – P. 303-314.

109. Экспоненциальное сглаживание [Электронный ресурс] : [Вэб-сайт]. – Электронные данные. – Режим доступа: http://www.machinelearning.ru. – Название с экрана.

110. Skulysh M. Hybrid resource management system for telecommunication network / M. Skulysh, S. Sulima // Advanced Information Systems. — 2018. — Vol. 2, № 1. — P. 47–51.

111. Суліма С. В. Система управління ресурсами в центрах обробки даних оператора мережі мобільного зв'язку / С. В. Суліма, М. А. Скулиш // Вісник НТУУ «КПІ». Серія Радіотехніка, Радіоапаратобудування. — 2017 . — № 68. — С. 27-32. (Наукове фахове видання України з технічних наук. Входить до міжнародних наукометричних баз: DOAJ, Web of Science, Index Copernicus, eLibrary.ru / РИНЦ та ін.)

112. Суліма С. В. Гібридна система управління ресурсами для віртуалізованих мережевих функцій / С. В. Суліма, М. А. Скулиш // Радіоелектроніка, інформатика, управління.— 2017 . — № 1(40). — С. 16–23. (Наукове фахове видання України з технічних наук. Входить до міжнародних наукометричних баз: Web of Science, DOI, DOAJ, British Library, eLibrary.ru / РИНЦ, Index Copernicus та ін.)

113. Суліма С. В. Алгоритм відображення та планування віртуалізованих функцій в мережі мобільного зв'язку / С. В. Суліма, М. А. Скулиш // Проблеми

телекомунікацій ПТ-2016 : 10-а міжнародна науково-технічна конференція, 19–22 квітня 2016 : матеріали конференції. — Київ, 2016. — С. 372–374.

114. Network Functions Virtualisation (NFV) Release 2; Management and Orchestration; Or-Vnfm Reference Point - Interface and Information Model Specification [Online]. – Available at: https://www.etsi.org/deliver/etsi_gs/NFV-IFA/001_099/005/02.04.01_60/gs_nfv-ifa005v020401p.pdf.

115. Network Functions Virtualisation (NFV) Release 2; Management and Orchestration; Or-Vi reference point - Interface and Information Model Specification [Online]. – https://www.etsi.org/deliver/etsi_gs/NFV-IFA/001_099/007/02.04.01_60/gs_nfv-ifa007v020401p.pdf.

116. 3GPP Specification TS: 23.107 Quality of Service (QoS) concept and architecture
https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=783

117. 3GPP Specification TS: 23.402 Architecture enhancements for non-3GPP accesses
https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=850

118. Руководство по технологиям объедененных сетей, 4-е издание.: Пер. с анг.: - М.: Издательский дом «Вильямс», 2005. – 1040 с.: ил.-Парал. тит. англ. ISBN 5-8459-0787-X.

119. M. Handley, H. Schulzrinne, E. Schooler, J. Rosenberg. SIP: Session Initiation Protocol.//RFC 2543, March 1999.

120. Tanir Ozcelebi, Igor Radovanovich. Multimedia Adaptation with SIP Resource Availability Signalling in IMS network.// IEEE 2007 International Conference on Convergence Information Technology. P. 1714-1719.

121. Скулиш М.А. Організація роботи групи серверів для забезпечення потреб розподіленої системи тарифікації послуг // Наукові записки Українського науково-дослідного інституту зв'язку. – 2014. – №5(33), 120-126сс.

122. Fernando A. Hernández Solana. QoS and Radio Resource Management in Multimedia Packet Transmission for 3G Wireless IP Networks./ Fernando A. Hernández Solana, A. Valdovinos Bardaji, Casadevall Palacio.// IEEE VTC 2004 Spring. – Milán, Italia.

123. Hao Wang. 4G wireless video communication, 1st edition./ John Wiley and Sons Ltd , 2009. P. 320.

124. Шилов Ф. І. Дослідження ефективності методу оптимального вибору обчислювальних ресурсів для білінгових систем / Ф. І. Шилов, М. А. Скулиш, А. Сафарян // Системи управління, навігації та зв'язку. - 2018. - Вип. 3. - С. 147-152.

125. Osadchuk, O. V., Semenov, A. O., Zviahin, O. S., Semenova, O. O., & Rudyk, A. V. (2021). Increasing the sensitivity of measurement of a moisture content in crude oil. Natsional'nyi Hirnychyi Universytet. Naukovyi Visnyk, (5), 49-53.

126. Кравчук, С. О., Лисенко, О. І., Явіся, В. С., & Новіков, В. І. (2021). Прикладні аспекти системного аналізу в телекомунікаціях та радіотехніці. Методичні рекомендації до виконання практичних занять.

127. Кравчук, С. О., Лисенко, О. І., Явіся, В. С., & Новіков, В. І. (2021). Основи теорії цифрових систем автоматичного керування: LTI моделі для систем SISO та MIMO.

128. Дослідження операцій [Електронний ресурс] : конспект лекцій / НТУУ «КПІ» ; уклад. О. І. Лисенко, І. В. Алєксєєва. – Електронні текстові дані (1 файл: 4,26 Мбайт). – Київ : НТУУ «КПІ», 2016. – 196 с. – Назва з екрана.

129. Semenova, O., Semenov, A., Voitsekhovska, O., & Kozin, D. (2020, October). The Neural Network for Vertical Handover Procedure. In 2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T) (pp. 753-756). IEEE.

Електронне мережне навчальне видання

М.А. Скулиш, С.В. Суліма

ОСОБЛИВОСТІ ОБЧИСЛЮВАЛЬНОЇ ІНФРАСТРУКТУРИ ДЛЯ СИСТЕМ КЕРУВАННЯ ТЕЛЕКОМУНІКАЦІЯМИ

Навчальний посібник

Реєстр. № 22/23-046 Обсяг 11 авт. арк.

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
проспект Перемоги, 37, м. Київ, 03056
https://kpi.ua
Свідоцтво про внесення до Державного реєстру видавців, виготовлювачів
і розповсюджувачів видавничої продукції ДК № 5354 від 25.05.2017 р.